43333

National Library of Canada
Bibliothèque nationale du Canada

Canadian Theses Division
Division des thèses canadiennes

Ottawa, Canada
K1A 0N4

PERMISSION TO MICROFILM — AUTORISATION DE MICROFILMER

• Please print or type — Écrire en lettres moulées ou dactylographier

Full Name of Author — Nom complet de l'auteur

Peter Frank Assmann

Date of Birth — Date de naissance

04/08/53

Country of Birth — Lieu de naissance

Netherlands

Permanent Address — Résidence fixe

9808 - 84 Ave, Edmonton, Alberta T6E 2G2

Title of Thesis — Titre de la thèse

The role of context in vowel perception

University — Université

Univ. of Alberta

Degree for which thesis was presented — Grade pour lequel cette thèse fut présentée

MSc

Year this degree conferred — Année d'obtention de ce grade

1979

Name of Supervisor — Nom du directeur de thèse

Dr. T. Nearey

Permission is hereby granted to the NATIONAL LIBRARY OF CANADA to microfilm this thesis and to lend or sell copies of the film.

The author reserves other publication rights, and neither the thesis nor extensive extracts from it may be printed or otherwise reproduced without the author's written permission.

L'autorisation est, par la présente, accordée à la BIBLIOTHÈQUE NATIONALE DU CANADA de microfilmer cette thèse et de prêter ou de vendre des exemplaires du film.

L'auteur se réserve les autres droits de publication; ni la thèse ni de longs extraits de celle-ci ne doivent être imprimés ou autrement reproduits sans l'autorisation écrite de l'auteur.

Date

Oct 17, 1979

Signature

Peter Assmann

NL-91 (4/77)

National Library of Canada

Cataloguing Branch
Canadian Theses Division

Ottawa, Canada
K1A 0N4

Bibliothèque nationale du Canada

Direction du catalogage
Division des thèses canadiennes

NOTICE

The quality of this microfiche is heavily dependent upon the quality of the original thesis submitted for microfilming. Every effort has been made to ensure the highest quality of reproduction possible.

If pages are missing, contact the university which granted the degree.

Some pages may have indistinct print especially if the original pages were typed with a poor typewriter ribbon or if the university sent us a poor photocopy.

Previously copyrighted materials (journal articles, published tests, etc.) are not filmed.

Reproduction in full or in part of this film is governed by the Canadian Copyright Act, R.S.C. 1970, c. C-30. Please read the authorization forms which accompany this thesis.

THIS DISSERTATION
HAS BEEN MICROFILMED
EXACTLY AS RECEIVED

AVIS

La qualité de cette microfiche dépend grandement de la qualité de la thèse soumise au microfilmage. Nous avons tout fait pour assurer une qualité supérieure de reproduction.

S'il manque des pages, veuillez communiquer avec l'université qui a conféré le grade.

La qualité d'impression de certaines pages peut laisser à désirer, surtout si les pages originales ont été dactylographiées à l'aide d'un ruban usé ou si l'université nous a fait parvenir une photocopie de mauvaise qualité.

Les documents qui font déjà l'objet d'un droit d'auteur (articles de revue, examens publiés, etc.) ne sont pas microfilmés.

La reproduction, même partielle, de ce microfilm est soumise à la Loi canadienne sur le droit d'auteur, SRC 1970, c. C-30. Veuillez prendre connaissance des formules d'autorisation qui accompagnent cette thèse.

LA THÈSE A ÉTÉ
MICROFILMÉE TELLE QUE
NOUS L'AVONS REÇUE

THE UNIVERSITY OF ALBERTA

The Role of Context in Vowel Perception

by

© P. F. Assmann

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES AND RESEARCH

IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE

OF Master of Science

IN

Speech Production and Perception

Department of Linguistics

EDMONTON, ALBERTA

Fall, 1979

THE UNIVERSITY OF ALBERTA

FACULTY OF GRADUATE STUDIES AND RESEARCH

The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies and Research, for acceptance, a thesis entitled "The Role of Context in Vowel Perception", submitted by Peter F. Assmann in partial fulfillment of the requirements for the degree of Master of Science in Speech Production and Perception.

.....
Supervisor

.....

.....

.....
External Examiner

Date ..... Oct. 17, 1979 .....

ABSTRACT

Vowel perception appears to be very robust: in the complete absence of any form of context vowels are recognized at a very high rate of identification (over 92 percent correct). This result is the outcome of a series of experiments which were conducted to assess the role of speaker differences and consonantal context in vowel perception.

Chapter one reviews previous studies which demonstrate a considerable amount of variability in vowel formant frequencies as a function of context. Since formant frequencies are generally believed to be the principal determinants of vowel quality differences, such variability may be expected to have an effect on perception. A number of hypotheses are discussed, and evidence for and against each one is presented.

In the second chapter, it is shown that labeling difficulties can have a profound effect on the outcome of vowel identification experiments, resulting in an over-estimation of the role of consonant context. Several solutions to this problem are suggested.

Chapter three describes two experiments which assess the perceptual role of speaker and consonant context. The results of these experiments demonstrate that vowels in

isolation are readily identified, and that the improvement afforded by context is rather minimal. It is proposed that vowels may contain additional information in the form of dynamic characteristics such as duration and diphthongization. Evidence for this view is presented in a study involving gated vowels, from which dynamic characteristics have been removed. Under such conditions identification errors increase and the role of speaker context is enhanced. Some possible implications of this finding for the perception of vowels in connected speech are discussed.

Chapter four presents acoustic measurements and detailed phonetic transcriptions for the data used in the gated vowel study. The statistical procedure of linear discriminant analysis is adopted to determine whether the identification responses of listeners can be predicted on the basis of information provided by acoustic measurements. Two specific questions are addressed: first, are tokens whose formant frequencies have values which are shared by more than one vowel category misidentified by listeners? Secondly, are tokens whose formant frequencies are characteristic of a single vowel category more likely to be identified correctly?

Under a speaker normalization hypothesis, vowel identification in the absence of speaker context (the "mixed" speaker condition) is determined primarily by the formant frequencies; when speaker context is available (the

"blocked" speaker condition), adjustments are made for speaker differences. To test this hypothesis, perceptual data from the mixed and blocked speaker conditions were compared with the output of discriminant analyses of raw formant frequencies and speaker-normalized formant values.

Significant correlations were obtained between the proportion of correct identifications of each vowel token and the probability of membership in the category intended (as determined by the classification algorithm of discriminant analysis). These findings indicate that vowel identification responses are closely related to formant frequencies. However, both raw and normalized formant values were more closely related to identification responses in the blocked than in the mixed condition. When speaker context is unavailable, formant values are not the only factors determining listeners' performance. Further experiments are needed to isolate these factors.

Phoneticians' judgements of height and advancement showed highly significant correlations with log F1 and log F2 values. A regression analysis indicated that the inclusion of speaker-dependent parameters significantly strengthens the relationship. Fundamental frequency and higher formants also make a significant contribution. These findings provide indirect support for the role of speaker information and vowel-internal cues in determining judgements of vowel quality.

Chapter five summarizes the major findings and arrives

at some conclusions concerning the perceptual significance of context. Further experiments are proposed to test hypotheses which are suggested by these results.

ACKNOWLEDGEMENTS

I would like to thank the members of my committee for their kind assistance and guidance in the preparation of this work.

I am particularly indebted to my thesis supervisor, Dr. T. Nearey, who was a continual source of ideas and inspiration at all stages of this project. His generous assistance, suggestions and guidance have been greatly appreciated.

I would also like to thank Dr. J. Hogan for his help, especially in statistical matters. I am grateful to Dr. A. Rozsypal and Dr. K. Holden for their careful readings of several versions of the thesis, and for their helpful comments and suggestions.

Special thanks go out to my fellow students and friends who provided moral support and helpful advice; especially my friend and colleague Ron Smyth.

This work is dedicated to my wife Albi.

Table of Contents

Chapter — Page

I. INTRODUCTION....1

Speaker Differences....3

Consonant Context....6

Hypotheses about the perceptual role of context....9

II. TASK DEMANDS AND VOWEL IDENTIFICATION....27

A. Experiment 1: spoken versus written responses....31

Listeners....31

Stimulus materials....32

Speakers....32

Apparatus....32

Procedure....36

Results and Discussion....39

Response frame and consonant context....45

B. Experiment 2: Keyword versus spelling responses....46

Listeners....47

Stimulus materials....48

Apparatus....48

Procedure....48

Results and Discussion....49

III. PERCEPTUAL STUDIES ON THE ROLE OF CONTEXT....56

A. Experiment 3: consonant context and vowel identification....57

Listeners....58

Stimulus materials.....58
Apparatus.....59
Procedure.....59
Results and Discussion.....60
Spelling and keyword responses.....60
IPA responses.....65
B. Experiment 4: Speaker context and vowel identification.....71
Listeners.....72
Stimulus materials.....72
Apparatus.....73
Procedure.....74
Results and Discussion.....74
IV. ACOUSTIC AND PERCEPTUAL STUDIES OF GATED VOWELS....82
A. Experiment 5: mixed and blocked speaker context and the identification of gated vowels..84
Listeners.....84
Stimulus materials.....84
Apparatus.....85
Procedure.....85
Results and Discussion.....86
B. Acoustic analyses of gated vowels.....92
Stimulus materials.....94
Apparatus.....94
Procedure.....94
Statistical analysis.....96
C. Phoneticians' transcriptions of gated vowels...111
Phoneticians.....112

Stimulus materials.....112
Apparatus.....113
Procedure.....113
Statistical analysis.....115
Correlations with acoustic data.....116
V. SUMMARY AND CONCLUSIONS.....122
REFERENCES.....125
APPENDIX 1.....133
APPENDIX 2.....134

LIST OF TABLES

TABLE.....PAGE

1. Experiment 1: vowel identification errors.....39
2. Examples of orthographically ambiguous words.....41
3. Experiment 1: ANOVA with simple main effects test.....42
4. Experiment 1: vowel confusion matrix.....44
5. Experiment 2: vowel identification errors.....51
6. Experiment 2: Partially repeated measures ANOVA with planned comparisons.....52
7. Experiment 2: vowel confusion matrix.....54
8. Experiment 3: vowel identification errors.....61
9. Experiment 3: Partially repeated measures ANOVA with planned comparisons.....63
10. Experiment 3: vowel confusion matrix.....64
11. Experiment 3: vowel confusion errors (IPA responses).....66
12. Experiment 3: Partially repeated measures ANOVA on IPA data.....68
13. Experiment 3: vowel confusion matrix.....69
14. Experiment 4: vowel identification errors.....75
15. Experiment 4: vowel confusion matrix.....77
16. Experiment 5: vowel identification errors.....88
17. Experiment 5: ANOVA.....89
18. Experiment 5: vowel confusion matrix.....90
19. Vowel formant frequencies (hertz).....96
20. Statistical discrimination of vowels.....102

21. Correlations (Pearson r) between P(G|x) scores and identification rates for individual tokens.....102
22. Correlations (Pearson r) between P(G|x) scores (based on discriminant analysis with adjusted prior probabilities) and identification rates for individual tokens.....109
23. Multiple regression: phonetic judgements with formant measures.....115
24. Phonetic judgements of vowel segments.....118

LIST OF FIGURES

Figure.....Page

1. Block diagram of stimulus preparation and presentation.....35
2. Frequencies of F1 and F2 (KHz) in log scale.....98
3. Frequencies of F2 and F3 (KHz) in log scale.....99

I. INTRODUCTION

Vowel formant frequencies are generally believed to be the principal determinants of vowel quality. The importance of formant frequencies in vowel production has been recognized for some time (Chiba and Kajiyama, 1941). The results of spectrographic analyses by Joos (1948) and Potter and Peterson (1948) led these writers to hypothesize a significant perceptual role for vowel formants as well.

Joos (1948) found that satisfactory approximations to the entire range of known vowel qualities could be obtained by means of speech synthesis using only two formants. This finding coincided with the discovery of a correlation between the frequencies of the first two formants and the classificatory features of height and advancement. Essner (1947) and Joos (1948) showed that plots of the first formant (F1) with the second formant (F2) from measurements of spoken vowels correspond fairly closely to the traditional vowel diagram.

Further corroboration of these results is found in a study by Delattre, Liberman, Cooper, and Gerstman (1952). Two sets of two-formant vowels were synthesized using the Pattern Playback. In set one, four F1 values were combined with a range of values of F2; in set two four F2 values were combined with a range of F1 values. They selected 235 tokens

which provided the closest approximations to each of the 16 cardinal vowels, and presented them in a series of listening tests to 11 phonetically trained listeners. Identification scores were generally very high. The authors concluded that F1 and F2 provide sufficient information for identifying the vowels. Subsequent experiments showed that F3 and higher formants contribute little to intelligibility. The authors did find, however, that two-formant approximations to vowels that ordinarily have F2 and F3 very close together (eg. the French vowels / I e ɛ / were identified more successfully when the second formant was higher than that of the spoken vowels.

More recently, Carlson, Granstrom and Fant (1970) investigated the feasibility of two-parameter models of vowel perception. They conducted an experiment in which Swedish listeners were asked to match four-formant synthetic vowels with two-formant approximations of these vowels. They found that two-formant approximations were sufficient for the identification of any of the Swedish vowels. Listeners tended to set the second formant of the synthetic two-formant vowels at frequencies in the vicinity of F2 for back vowels, intermediate between F2 and F3 for open front vowels, and closer to F3 or perhaps higher for high unrounded front vowels. Carlson et. al. suggest that the "effective" second formant, or F2', is in the vicinity of the second formant of the two formant stimuli.

These results are generally taken as evidence that a

single frequency-amplitude section of a vowel which takes into account the frequency of F1 and F2 (and possibly F3) contains sufficient information for the identification of the vowel. However, it is known that listeners make use of other available information in naturally spoken vowels. Such factors include duration and fundamental frequency. Duration differences may help to distinguish vowels which are similar in their formant frequencies (Tiffany, 1959). Miller (1953) found that when the fundamental frequency of synthetic two-formant vowels was doubled, listeners showed shifts in their categorization of certain vowels. The tokens most affected were those occupying the boundary areas of /ʊ/, /ʌ/, and /æ/.

While it appears that such factors may contribute to the perception of vowel quality differences, they are usually regarded as supplemental or secondary cues, and seem to be quite variable in natural data. A speaker may change the pitch of his voice, for example, or alter durations by changing his speaking rate.

Speaker Differences

It has been known for some time that the absolute values of the frequencies of F1 and F2 show considerable variation across speakers (Joos, 1948; Potter and Peterson, 1948). The first systematic large-scale study of vowel formant frequencies was undertaken by Peterson and Barney (1953). They recorded 10 vowels in / hVd / frame from 76

speakers of "General American" (33 men, 28 women, and 15 children). Measurements of f0, F1, F2, and F3 frequency were taken from spectrograms. Plots of F1 by F2 for all speakers resulted in considerable overlap between certain vowels, particularly those which occupy the area near the centre of the F1-F2 plane. Inclusion of f0 and F3 did little to reduce overlap, except in the case of the vowel / ɝ /.

In part, these differences reflect vocal tract size differences between the men, women and children who served as speakers. On the average, F1 and F2 are higher for women than for men, while children have the highest formant range.

An additional source of overlap in formant frequencies may lie in dialect differences. Of the 76 speakers, most of the women and children grew up in the Mid-Atlantic region of the United States while the men came from a "broader regional sampling" of the U.S. (Peterson and Barney, 1952, p. 177). Two speakers were born outside the U.S., while a few others spoke English only as a second language.

According to traditional phonetic theory, the vowels of different speakers of the same dialect can be equated in terms of vowel quality. When phoneticians make such comparisons, they regard differences between the speakers of the same dialect as personal characteristics which do not enter into judgements of vowel quality (Ladefoged, 1967).

In order to make meaningful comparisons (eg. in terms of acoustic measurements) between different tokens of the "same" vowel it seems reasonable to restrict investigation

to a single dialect, as judged by one or more competent phoneticians. While a complete theory of speech perception must eventually account for a listener's ability to understand speakers from different dialect areas or speakers with "foreign accents", there are indications that listeners may rely heavily on external information (such as semantic context) to decode their speech. If information of this sort is not available, intelligibility may be impaired.

In conjunction with their measurement study Peterson and Barney also conducted listening tests. These tests made use of the same vowels which were analyzed acoustically. Of the 70 listeners, 32 also served as speakers in the measurement study. The vowels / ɑ / and / ɔ / were frequently misidentified, which may reflect that certain U.S. dialects do not differentiate these vowels.

Plots of F1 and F2 for vowels which received 100 percent correct recognition scores showed considerably less overlap in the vowel areas. However, a substantial amount of residual overlap was still observed.

Fairbanks and Grubb (1961) found a strong correlation between the degree of formant overlap and listeners' judgements of "representativeness" on a scale of 1 to 9. In spite of this correlation, some tokens which were judged to be highly representative of the vowel intended showed overlap with other vowels in their formant frequencies.

Measurement studies of the vowels of other languages have confirmed this basic finding: there is a considerable

amount of variability in the frequencies of F1 and F2 across speakers, resulting in overlap between a number of different vowel pairs. Studies have been conducted on the vowels of Swedish (Fant, 1959), Japanese (Fujisaki and Kawashima, 1967), Dutch (Pols, Tromp and Plomp, 1972) and a number of other languages.

The extent of this variability is considerable; in the Peterson and Barney data, differences may be as large as 350 Hz (35 percent) in F1, and 900 Hz (28 percent) for F2. As Joos (1948) pointed out, if such discrepancies existed within the speech of a single individual, the phonetic distinctness of the vowels would be lost.

Consonant Context

A number of investigations have suggested that the formant frequencies of vowels undergo modification in different contexts of stress, rate and consonant environment. The present discussion will be limited to studies of vowels in stressed syllables spoken at normal rates; ie. in citation forms.

Variations in the formant frequencies of vowels throughout the course of a word or syllable were pointed out by Potter and Steinberg (1950). Potter, Kopp and Green (1950) refer to changes associated with the onset and offset of a vowel as initial and final "influences". Formant frequency measurements were consequently taken by these authors at the steady state portion of the vowel, where such

variations are at a minimum.

More recent investigations have shown that vowel formant frequencies may differ as a function of the phonetic context. Stevens and House (1963) measured F1, F2 and F3 values for a set of eight vowels in different consonant contexts. Three speakers repeated bisyllabic nonsense words, / hə CVC / (where V represents one of the 8 vowels and C represents one of 14 consonants; both initial and final C's refer to the same consonant. Analysis-by-synthesis (spectrum-matching) procedures were used to generate formant measurements. These were taken at three equidistant points from the onset to the offset of the vowel, and were averaged to obtain a single set of measures for each vowel.

When the vowels were spoken in isolation or in / hVd / context (subsequently termed "null" contexts by the authors), formant measurements were in general agreement with those of Peterson and Barney (1952). However, comparing vowels in the null environment with those in other CVC contexts, they found systematic deviations in formant frequencies, in either a positive or negative direction. F1 differences were generally quite small. Deviations in F2 were always toward a more "central" position: for front vowels with a high F2, a downward shift was observed (with larger shifts for short vowels); for back vowels with low F2 the shift was upwards. Front vowels were affected most by labial and alveolar environments, least by velars; for the back vowels the labials caused only small shifts while

alveolars resulted in large deviations.

Stevens and House suggest that formant displacement may be explained in terms of the dynamic properties of the articulatory mechanism. The amount of displacement of F2 is dependent on the distance traveled by the articulators from the initial position of the first consonant, to the target position of the vowel, to the terminal position of the final consonant. In acoustic terms, formant frequency shifts are from the "target" frequency of the vowel (ie. from its value in the "null" environment) toward the formant loci of the adjacent consonants.

Lindblom (1963), using a single speaker, measured formant frequencies and durations for 8 Swedish vowels in CVC frames under different conditions of stress (stressed vs. unstressed), consonant context (/ bVb /, / dVd /, / gVg /), and position in the carrier sentence (initial vs. final). F2 values were found to differ from their values in the null environment in all CVC contexts. These deviations were generally in the direction of F2 for the adjacent consonants. Lindblom proposed a model to describe formant values as an exponential function of duration. Generalizing from this model, he proposed that vowel "targets" underlie a speaker's intended utterances, but due to mechanical constraints on the articulators, formant values actually obtained fall short of these values (an "undershoot" effect). Thus differences in formant frequencies are attributed to time limitations on the attainment of vocal

tract target positions, rather than to stress, position or consonant context *per se*. Stressed syllables show less variation in formant values than unstressed syllables; but as predicted by the model, they are also longer in duration.

The universality of the undershoot model is questioned by the findings of Stevens and House (1963:124) of considerable variation among speakers in the degree of deviation in F2 for different consonant environments. With a larger number of speakers even greater heterogeneity is to be expected.

Hypotheses about the perceptual role of context

Several kinds of hypotheses have been advanced to describe the ways in which listeners may deal with the lack of invariance in vowel formant frequencies across speakers and in different consonant environments (see Nearey, Hogan and Rozsypal (1979)).

1. Variations may be subthreshold, or undetectable by listeners. On this hypothesis, context differences simply do not play a role in vowel perception; they represent random variations induced by constraints on the production process. An important corollary of this hypothesis seems to be that formant frequency variations will not result in overlap of vowel areas.
2. A second hypothesis maintains that contextual variability *can* be detected, but is actively "tuned out"

or ignored by listeners. Formant frequency variations are not predictable on the basis of context, but constitute perceptual "noise" for the listener. Whenever formant frequency variations result in overlap in vowel areas, the listener relies on information from other sources in the signal to disambiguate the vowels. These sources may include f0 and higher formants, duration, and diphthongization. On this view, invariant information for the identification of a vowel may come from other "vowel internal" sources.

Another possibility is that listeners attend to these various sources simultaneously, basing their decisions on some weighted combination of the available parameters.

3. The third hypothesis also holds that variations are detectable, and moreover, that these variations are systematic and are predictable on the basis of the contexts in which they occur. The listener attends to these variations and "normalizes", or adjusts his perceptual criteria to recover the phonetic identity of tokens whose formant frequencies in isolation might associate them with the wrong vowel class.

4. Finally, context may be <u>essential</u> for the successful recognition of the vowel; in other words, the vowel is <u>specified by its context</u>. This hypothesis predicts that context-free vowels will be difficult or impossible to identify.

Each of these hypotheses offers an account of the perceptual role of speaker or consonant context differences. They are not mutually exclusive; it is possible that more than one hypothesis is needed to account for the results obtained under different experimental conditions. Note that the difference between the third and fourth hypothesis may be a matter of degree; the latter may be viewed as a limiting case of the former. Speaker context and consonant context may function in different ways, requiring separate accounts of their role in vowel perception.

In order to test the hypothesis that formant frequency deviations lie outside the detection threshold, synthetic speech is used to systematically control and manipulate the relevant parameters. Previous studies of just noticeable differences (JND's) in formant frequencies have generally sampled only selected small portions of the relevant space.

Synthetic steady state vowels were generated by Flanagan (1955b). Setting F1 and F2 at several representative values, he changed their values in steps of 10, 20, 30, 40, 50, 60, and 70 Hz. He obtained difference limens (DL's) of approximately three to five percent of the formant frequency.

Flanagan (1972) suggested that this experiment be extended to cover the entire F1-F2-F3 space. In real speech formants vary simultaneously; in his experiment only one formant was varied, holding the others constant. In addition, vowel formant frequencies in CVC context show

variations throughout the course of the syllable.

The effects of consonant transitions on difference limens for formant frequency were assessed by Mermelstein (1978). DL's were measured for two steady state vowels and the same vowels embedded in symmetrical voiced stop consonant environments. The contexts chosen were / bVb / and / gVg /, which differ only in the trajectory of F2 transitions. DL's for time-varying CVC stimuli were found to be consistently larger than those in steady state vowels, indicating that formant frequency differences are less well discriminated in CVC context. For vowels in the / ɛ - æ / boundary region, DL's for F1 averaged 33 Hz in steady state vowels (about 6 percent of F1), but 70 Hz in CVC context (about 12 percent of F1); average DL's for F2 were 75 Hz in steady state vowels (about 4 percent of F2), and 171 Hz in context (about 10 percent of F2). In each case DL's were twice as large in CVC context, greater than 10 percent of the formant frequency.

On the average, changes of 60 Hz in F1 and 176 Hz in F2 were not reliably discriminated by listeners. The average shift in F1 across all consonants, as reported by Stevens and House (1963:p.117) approaches this value only in the case of / ʌ /; average shifts in F2 exceed these limits in only two cases, for the vowels / I / and / u /. Mermelstein questions the perceptual significance of these shifts. In order to have a consistent effect on vowel identification they must be reliably discriminable.

Speaker differences, on the other hand, consistently exceed the difference limens in both F1 and F2. Average differences in F1 between adult males and children range from 70 Hz for / ɜ / to 350 Hz for / æ /. Average F2 differences are even larger, ranging from 220 Hz for / ɔ / to 910 Hz for / i /.

Circumstantial evidence for the second hypothesis comes from a study by Ainsworth (1972a) using synthetic speech. When synthetic vowels were supplied with "natural" durations, standard deviations for categorization decreased and boundaries between vowel areas showed greater stability. A similar finding was reported by Miller (1953) when F3 was introduced into synthetic two-formant vowels. These findings suggest that other vowel-internal parameters help to specify the identity of the vowel, especially in boundary regions where the formant pattern may characterize more than one vowel category.

Hypothesis two predicts that the vowel identification process will not be affected, one way or the other, by the addition of context. Listeners are able to compensate for contextual variation in formant frequency by reference to additional vowel parameters. Context-free vowels should be well-identified.

The third hypothesis attaches a special role to context, and several versions of this hypothesis have been elaborated.

Potter and Steinberg (1950) point out that a vowel in

the context of a CVC syllable, undergoes transitional movements from initial to final consonant. They suggest that such changes may aid in the identification process. Peterson and Barney (1953:p.184) appear to share this view, stating that "words are not adequately represented by a single section, but require a more complex portrayal". Mermelstein, Liberman and Fowler (1978) point out that "consonant" transitions and "vowel" steady states are not readily differentiated in natural speech; they consider it unlikely that decisions concerning the the phonetic identity of the consonant and the vowel are made independently by the listener.

Lindblom and Studdert-Kennedy (1967) elaborate a model for vowel perception in which consonant context is taken into account. They propose that:

> Other information in the short-term acoustic context, such as the direction and rate of adjacent formant transitions, may also be important in the auditory representation of the signal and the process of symbol assignment. (p. 832)

According to their model, listeners compensate for formant frequency differences in consonant frames, which are the result of an "undershoot" effect in production. "Perceptual overshoot" is a mechanism which alters the categorization of a vowel in accordance with expectations generated by the context in which it is found.

Lindblom and Studdert-Kennedy describe an experiment in which English listeners were presented with synthetic vowels on a continuum between / I / and / U /. They compared the

categorization of the vowels in isolation and in two consonant frames, / jVj / and / wVw /. Results indicated that the categorization of the vowel shifted as a function of the direction and rate of adjacent formant transitions. Shifts were in the direction predicted by a hypothesis of "perceptual overshoot", but were smaller than anticipated in the / jVj / context.

Since the two consonant frames used in the Lindblom and Studdert-Kennedy study violate English phonotactic constraints, it is important to know whether similar results could be obtained in other consonant environments. The proposed model of perceptual overshoot is a version of the third hypothesis: listeners are able to compensate for variations in vowel formant frequencies by reference to information in adjacent consonants. This hypothesis does not predict that vowels pronounced in isolation will be difficult to identify, although vowels excised from CVC context may be.

A number of hypotheses have been proposed to describe the perceptual role of context provided by individual speaker characteristics. One of the first was put forward by Joos (1948). Joos suggested that listeners are able to compensate for variations in formant frequencies by observing the range of variation present in the vowels of a given individual. This range serves as the basis for a "coordinate system" against which further vowels are compared for their identification.

Invariant information is thus to be found in the relationships present among the formant frequencies of the vowels of a single speaker; these relationships must be determined separately for each speaker on the basis of formant ranges specified by other vowels.

Several versions of the speaker normalization hypothesis have been proposed. They attempt to describe the adjustments which might be required to compensate for individual speaker characteristics. Ladefoged (1967) suggests that listeners may scale their responses to a weighted mean of the formant frequencies of vowels present in prior speech. Previous tokens of the vowels serve as an anchor or standard, creating an internal adaptation level to which successive vowels are refered. This hypothesis is based on Helson's (1948) "adaptation level" theory which attempts to account for perceptual constancies in a number of sensory modalities.

Ladefoged and Broadbent (1957) conducted an experiment to determine whether normalization shifts could be induced by presentation of different kinds of contextual material. They compared listeners' categorizations of synthetic four-formant / bVt / syllables, using the vowels / I, ɛ , æ, ʌ, /. The context sentence was 'Please say what this word is: .....'. Six versions of the sentence were generated by raising or lowering either F1 or F2. The differences in context sentences were apparently interpreted as changes in speaker quality by the listeners. Shifts were

obtained for each of the four test syllables in different sentence contexts. Changes in categorization were in the direction predicted by a formant normalization hypothesis.

Lieberman (1973) and Lieberman, Crelin and Klatt (1972) have suggested that the "point" vowels / i, a, u / may play a special role in calibrating the vowel space of a speaker. These vowels occupy the extremes of the acoustic and articulatory vowel space. They are more acoustically stable than other vowels with respect to perturbations in articulation, and they are the only vowels whose formant frequencies can be specified by unique area functions. Gerstman (1967) developed a normalizing algorithm which successfully classified the vowels of the Peterson and Barney corpus (97 percent correct classification). The algorithm performs a linear rescaling of F1 and F2 based on the position of each formant in the frequency range of a given subject. This range is calculated from the endpoint values, positions occupied by the vowels / i /, / a / and / u /.

Nearey (1977) describes two types of relative formant normalization procedures. Range normalization procedures require that the formant frequencies from two or more vowels of known phonetic quality be specified in advance. All of the models described above would fall under this category.

The second type, called point normalization procedures, depend on the prior specification of formant values from a single vowel of known phonetic quality. One version of this

hypothesis, the constant ratio hypothesis, is based on the observation that the ratios of the formant frequencies of the vowels appear to be constant over a wide range of speakers, including men, women and children. Formant values for a given speaker's vowels can therefore be estimated on the basis of a single speaker-dependent scale factor. This scale factor could be derived from the formant values of a single vowel whose phonetic identity is known. Nearey suggests that the vowel / i / may be especially well-suited as a basis for perceptual normalization of the sort described by this hypothesis. This vowel is the least ambiguous in terms of formant overlap. Perceptually it is well-recognized under a variety of circumstances (Peterson and Barney, 1952; Tiffany, 1953).

In a perceptual study using synthetic speech, Nearey found that all of the phonetic boundaries of a set of synthetic two-formant vowels were shifted as a function of changes in the formant frequencies of a single context vowel / i /. The formant frequencies in the context vowel were altered to give the impression of a change in speaker identity (formant ranges were characteristic of an adult male and a child), without affecting listeners' judgements of the phonetic identity of the vowel (in each case the test vowel was classified as an / i /). Upward boundary shifts in F1 and F2 were found for each vowel as the context vowel was changed from a "male's" to a "child's" / i /. Shifts in the categorization of test vowels were in the direction

predicted by the constant ratio hypothesis.

Few studies have investigated the normalization hypothesis with natural speech. Normalization shifts have all been demonstrated with synthetic speech, indicating that listeners may be able to use the information inherent in the formant relationships of a single speaker's vowels. In natural speech it is not possible to manipulate formant frequency values precisely; therefore categorization shifts in vowel boundaries are difficult to map.

Dechovitz (1977) offers one solution to the problem. If the relationships between context vowels spoken by a given speaker can affect a listener's categorization of a target vowel, it should be possible to induce misclassifications by mismatching speaker identity between the context vowels and the target vowel. Misclassifications will occur when the formant frequencies of context vowels are incompatible with the intended phonetic identity of the target. Such errors will occur only when the formant frequencies of the target vowel overlap between two alternative choices, and disambiguating sources of information are not available.

Dechovitz used naturally spoken tokens of / bVb / syllables embedded within a carrier sentence frame with five vowels of American English. Two speakers were recorded: one was an adult male, the other a nine-year old child. The adult speaker mimicked the speaking rate and pitch level of the child's utterances. When test syllables spoken by the adult male were excised from the sentence and embedded

within the child's carrier, an increase in identification errors was found compared with the identification of the syllables in isolation or in the appropriate carrier frame. Acoustic analyses of the vowel confusion errors (/ bɛt / - / bʌt / and / bæt / - / bɑt /) indicated that these vowels overlap in their formant frequencies between the two speakers. The results are taken as evidence for the view that listeners can adjust their criteria for vowel identification in accordance with formant ranges specified by the context.

Strange, Verbrugge, Shankweiler and Edman (1976) and Verbrugge, Strange, Shankweiler and Edman (1976) conducted a series of studies to determine the effects of speaker differences and consonant context on vowel identification. They predicted that uncertainty about the speaker, as induced by a speaker randomized condition (MIXED speaker condition) would result in more confusion errors than a series in which the speaker was fixed (BLOCKED speaker condition). If extended exposure to vowels from a single speaker is necessary to allow normalization to take effect, the improvement in the blocked condition should be considerable. The number of errors in the mixed speaker condition should, on a formant normalization hypothesis, reflect the degree of overlap in formant frequencies.

Similarly, if consonant context provides essential vowel information, CVC syllables should fare better than isolated vowels.

In their initial study 15 vowels, including 5 diphthongs, were recorded in / hVd / frame from 30 speakers including 13 men, 12 women, and 5 children. Vowels were presented in the mixed speaker condition and preceded by three precursor syllables, / kip /, / kap / and / kup /, pronounced by the same speaker. Verbrugge et. al. reasoned that if exposure to the point vowels enables a listener to calibrate a speaker's vowel space, the precursor condition should show a smaller error rate than the mixed speaker condition.

Results did not indicate the predicted improvement: 12.9 percent errors were made in the mixed condition, 12.2 in the precursor condition. Errors actually increased for some vowels, particularly those which are adjacent in the vowel space to the precursors (eg. / I , ɔ, U /. It is possible that some kind of context effect was operative (see Fry, Abramson, Eimas and Liberman, 1960). Verbrugge et. al. noted the presence of response biases in the precursor condition.

These findings are taken as evidence against a speaker normalization hypothesis. However, there are conditions under which a formant normalization hypothesis would anticipate results of this kind; for example, if the vowel categories were not aligned between speakers and listeners. If listeners and speakers differ in dialect, a normalization procedure may tend to misclassify vowels with atypical formant frequencies for the dialect in question. Recent

studies (Hindle, 1978) have suggested that the application of normalization procedures to the vowels of speakers from different dialects may actually help to reveal such differences. The 25 percent increase in / ɔ - ɑ / confusions in the precursor condition may be attributable to a mismatch in vowel categories. A large proportion of the listeners and speakers were native to the Upper Midwest where an / ɔ - ɑ / distinction does not occur.

The error rate in the mixed speaker condition was 12.9 percent. This figure is somewhat higher than 5.6 percent, obtained in a comparable experiment by Peterson and Barney (1953). However, the Peterson and Barney experiment was restricted to monophthongs, while the former experiment also included diphthongs. This difference cannot account for the increase in errors, since diphthongs are not frequently misidentified in the Verbrugge et. al. study. A further difference lies in the number of tokens recorded from each speaker, and the number of speakers: Peterson and Barney used twenty tokens from each of the 70 speakers.

In a further study, Strange et. al. (1976) recorded nine vowels in isolation and in / pVp / frame from five men, five women and five children. The speakers were required to read the test syllables aloud as written on cards. For the isolated vowels and for those / pVp / syllables which could not be rendered unambiguously into English spelling, a set of keywords was used (eg. 'vowel as in cawed').

Listeners heard vowels in either mixed or blocked

speaker condition, in / pVp / context or in isolation. The blocked speaker condition presented five tokens of each of the vowels from three speakers: (a man, a woman and a child) whose vowels were shown to be "representative" in terms of the numbers of errors each generated in the mixed condition. In the mixed speaker condition each of fifteen speakers contributed two repetitions of 3 of the nine vowels.

The same procedure was used for the isolated vowels and the / pVp / syllables. Separate groups of listeners were used in each of the four groups.

Results indicated a large increase in errors in the isolated vowel condition (42.6 percent in the mixed condition, 31.2 in the blocked) as compared with the / pVp / condition (17.0 percent in the mixed condition, 9.5 in the blocked). Both kinds of context showed statistically significant improvements, but the effects of consonant context were considerably greater. The absence of an interaction between the two factors was taken as an indication that consonant environment does not aid the perception process by providing cues for speaker normalization. The authors concluded that consonant context is critical for vowel identification, while speaker context is of marginal importance. Vowels are much more poorly identifed when they are presented in isolation than in consonant context.

Analysis of the acoustic parameters of the tokens used in this study indicated that vowel errors could not be

attributed to errors in production. Average formant frequencies and durations for men, women, and children correspond closely to the values obtained by Peterson and Barney (1953) and Peterson and Lehiste (1960). Vowels which were frequently misidentified were not necessarily aberrant in their formant frequency values. While some vowels were apparently "misarticulated", deviations in formant frequencies were as likely to occur in / pVp / syllables as in isolation. The authors concluded that acoustic information for vowel identity is "specified in the dynamic configuration of the syllabic pattern as a whole" (Strange et. al., 1976, p.219).

A subsequent experiment showed that listeners were able to identify CVC syllables with considerable accuracy even when both consonant frame and speaker were varied from token to token. Consonant frames consisted of CVC syllables, where 'C' refers to one of the six consonants / p, t, k, b, d, g / and 'V' refers to one of nine vowels used in the previous study. Consonant frames consisted of both symmetrical (eg. / bVb /) and non-symmetrical (eg. / bVd /) contexts. Thirty speakers were selected to maximize diversity in voice quality and vocal tract size. Even under these conditions, listeners misidentified only 22 percent of the vowels, compared with 42.6 percent for the comparable isolated vowel condition. The authors concluded that:

> The acoustic effects of coarticulation carry substantial information about a medial vowel, which aids in vowel identification whether or not the listener has prior knowledge of the consonant's

identity (Strange et. al., 1976, P.221)

High error rates for isolated vowels were taken to indicate that such "quasi-steady-states" are not well-formed utterances from the standpoint of the perceiver. Acoustic information that is generally regarded as belonging to the consonant is hypothesized to aid in the specification of adjacent vowels. In a subsequent paper (Shankweiler, Strange and Verbrugge, 1977) it is suggested that formant transitions aid in vowel identification and that information for the vowel is not localized, but is distributed throughout the entire syllable. The authors propose that consonant context is essential for the successful recognition of vowels.

If consonant context specifies the identity of the vowel as Strange et. al. (1977) have proposed, listeners should find it extremely difficult to identify vowels in isolation. Recent studies have challenged this conclusion.

Kahn (1978) conducted tests with 8 vowels in / hV / frame from 20 speakers in a mixed speaker context. Productions of the vowels were carefully monitored, and any aberrant tokens were discarded. An optimally simple answer sheet was prepared to minimize the difficulty of the task for listeners. The answer frames were actual words of English, and were unambiguous in their spelling. Ten phonetically trained listeners were tested. Each vowel was presented twice, to eliminate the effects of momentary lapses in attention.

Fewer than three percent errors were made by the listeners. There is some indication that some of the residual errors were attributable to dialect differences between speakers and listeners.

Kahn's results suggest that CV syllables which do not contain formant transitions are well recognized under the conditions investigated. This result conflicts with the findings of Strange et. al. Differences in the task, in the recording procedure and the screening of the tokens, and in the dialects of the speakers and listeners, may account for this discrepancy. Kahn's findings indicate that vowels are not necessarily ambiguous, even in the absence of consonant transitions and speaker context.

Kahn's study used only subjects who had received training in phonetics; the tokens used were carefully screened to eliminate mispronunciations. Moreover, listeners were given two opportunites to listen to the vowels; more if necessary. The use of an unambiguous response list may have simplified the task for listeners. One or more of these factors may have been responsible for the large differences in error rates between the two studies.

In the present study, a series of experiments was conducted to investigate the reasons for this discrepancy.

II. TASK DEMANDS AND VOWEL IDENTIFICATION

It has generally been assumed in vowel identification experiments that the number and nature of identification errors are a direct reflection of the acoustic properties of the signal (including its "context"). In particular, many studies have assumed that task demands are constant across a wide range of experimental conditions. The experiments described below were designed to compare the effects of different response types on the identification of vowels. If context plays a significant role in vowel perception, its effects on vowel identification should be constant across a variety of response conditions.

The first factor to be considered is orthographic interference. When the stimuli to be identified are actual words of English, one can simply require listeners to respond by marking off the words on an answer sheet which provides their spellings. However, it is often desirable to maintain a single consonant frame across a large set of vowels. There does not appear to be any single CVC frame which generates English words for all possible medial vowels. This entails the use of "nonsense words", for which a new spelling must be invented on analogy with similar "real" words. Secondly, in experiments with isolated vowels spelling responses are out of the question, since there is

no unique orthographic representation for many of the English vowels. More than one spelling form exists for most of the vowels, and most of the spelling forms can represent more than one vowel sound. For example, the letter 'u' may be pronounced as / ʊ / as in 'put', or as / ʌ / as in 'putt'; or as / u / as in 'flu'. The letters 'oo' can represent / ʊ / as in 'book', / ʌ / as in 'flood', or / u / as in 'boot'.

When spelling responses are inappropriate or impossible to use, one may require listeners to give "keyword" responses: answering with another word which contains the same vowel. When keyword responses are called for, the orthographic similarity of a given response to an alternative choice may make additional processing demands on the listener. This in turn may result in an increase in response latencies and an increase in the probability of selecting an incorrect alternative. The listener must learn to associate an unfamiliar stimulus (a vowel in isolation or in consonant frame) with one of a set of response alternatives (spelling forms which may or may not correspond to real words). Clearly this is a more complex task than simply matching a word with its spelling.

In the study conducted by Strange et. al. (1976) vowels were presented in isolation and in / pVp / syllables to a group of listeners, who were required to mark off their responses on an answer sheet containing the following alternatives:

pip pup pap peep pep pop poop pawp puup

The listeners were therefore required to give spelling responses to some of the stimuli and keyword responses to others. Since there are no English words / pep / or / pUp /, these "nonsense" words were given the most plausible spelling representations. In the case of / pUp /, however, there is no unique orthographic representation for the vowel: the investigators were forced to invent a new spelling, 'puup', which bears no resemblance to any existing English word. Listeners had to learn this new spelling, and it was necessary to draw special attention to the form and its pronunciation (Strange et. al. 1976, p. 215). It is not unreasonable to expect that some listeners might forget the new spelling, selecting instead the alternative 'pup' by analogy with English words like 'put'; or 'poop', by analogy with English words like 'hood' or 'wool'. One of the highest error rates in the Strange et. al. (1976, p. 216) study was obtained for the vowel / U /, the vowel represented in 'puup' In fact, when / ɑ - ɔ / confusions are eliminated (for reasons of dialect: see above) the vowel / U / far exceeds any other vowel in the proportion of errors. Many of these errors involve substitution of / ʌ / ('pup') for / U / ('puup').

In a footnote, Strange et. al. (1976, p. 223) describe an experiment in which the response frames were changed to correspond with the presumed spelling of the isolated

vowels: 'EE', 'IH', 'EH', etc. They report no improvement for the isolated vowels, but an increase in errors for the / pVp / syllables. Since isolated vowels do not occur as English words, spelling uncertainty may pose special problems for subjects attempting such a task: in effect, listeners must learn to use a new spelling or transcription system with partial correspondence to existing rules of English spelling. Further details concerning the results of this study are unfortunately not given.

Turning to the results for isolated vowels, it appears that orthographic interference may be present here as well. Strange et. al. obtained very high error rates for / ʌ / and / U /, both of which are rendered by a number of different spellings in English words like 'rough', 'put', 'could', 'pup', 'wood', 'full', 'flood', etc. Spelling confusions will obviously introduce a bias against isolated vowels.

A second problem concerns the labeling of unfamiliar stimuli. When stimuli are actual words of English, familiarity with the words may simplify the labeling task. In part, therefore, the large differences observed by Strange et. al. between CVC's and isolated vowels may be due to labeling difficulties rather than perceptual difficulties per se. Strange et. al. report that 38 of the 72 CVC's used in their experiment with variable consonant frames were actual English words. This subset showed an average error rate of 15 percent, compared with 25 percent for non-words. However, these results must be interpreted with caution,

since the number of English words is not constant across the entire set of vowels.

A. Experiment 1: spoken versus written responses

The present study was designed to assess the magnitude of non-perceptual task difficulties. Orthographic interference is a potential problem, whether one uses keyword or spelling responses. However, a third form of response circumvents the orthography problem: simply have listeners repeat what they hear, and have a panel of phonetically trained listeners transcribe their repetitions in terms of a standard transcription system.

The study conducted by Strange et. al. (1976) used response frames which required spelling responses for most of the / pVp / stimuli, but keyword responses for the isolated vowels. In order to balance the task difficulty equally across both contexts, the experiment to be described here combined repetition (spoken) responses with keyword responses, for both stimulus types.

Listeners

Listening tests were conducted in the language laboratory of an Edmonton high school. Listeners were 26 Grade 10 students of fairly homogeneous language and dialect

background. Prior to the experiment, each was asked to fill out a questionnaire concerning their language background, hearing status, etc. (see Appendix 1).

Stimulus materials

Isolated vowels and vowels in / pVp / frame were recorded. Ten monophthongal vowels were used in each set:

/ i, ɪ, e, ɛ, æ, ʌ, ɒ, o, ʊ, u, /

The vowel / ɜ / was excluded. The vowel / ɒ / represents a phonemic merger of / ɑ - ɔ / in Canadian English (Avis, 1975).

Speakers

To ensure homogeneity of dialect the speakers were matched as closely as possible for their language background. Five adult males and five adult females were selected. All of the speakers were residents of the Edmonton area, and in addition were residents of this city throughout most of their childhood and adult years.

Apparatus

The instruments used in this experiment are described below, along with their technical specifications:

1. Minicomputer: DEC PDP-12A. (Word length 12 bits; A/D, D/A converters 10 bits; operating systems OS/8 and

Alligator).[1]

2. Microphone: Sennheiser MD 421N. (Frequency response: 30 to 17000 Hz ±5 dB, with 5 dB rise between 3 and 10 KHz. Sensitivity: 0.2 mv/microbar for 1000 Hz.)
3. Tape recorder: (a) TEAC A-7030. (Frequency response: 50 to 15000 Hz ± 2 dB; speed 7.5 ips; S/N ratio: 58 dB.) (b) SONY portable tape recorder: TC-105 model.
4. Headphones: SONY DR-9.
5. Audio-frequency filter: Rockland 1524-01 (slope of frequency response: 30 to 6000 Hz ± 3 dB)

Recording

Speakers were recorded individually in a sound-treated room. The stimuli were recorded using the left channel of the TEAC, in order to avoid possible crosstalk effects.

In order to eliminate the possibility of orthographic interference speakers were given spoken "models" or "prompting stimuli". Prompting stimuli were recorded on a master tape using the vowels of the first speaker. This speaker had received phonetic training. He was given a list of phonetic symbols, and his productions were carefully monitored to avoid obvious mispronunciations.

Speakers were asked to repeat (rather than "mimic") the vowels they heard in a natural way. They were also given a transcript containing English equivalents for each of the

---

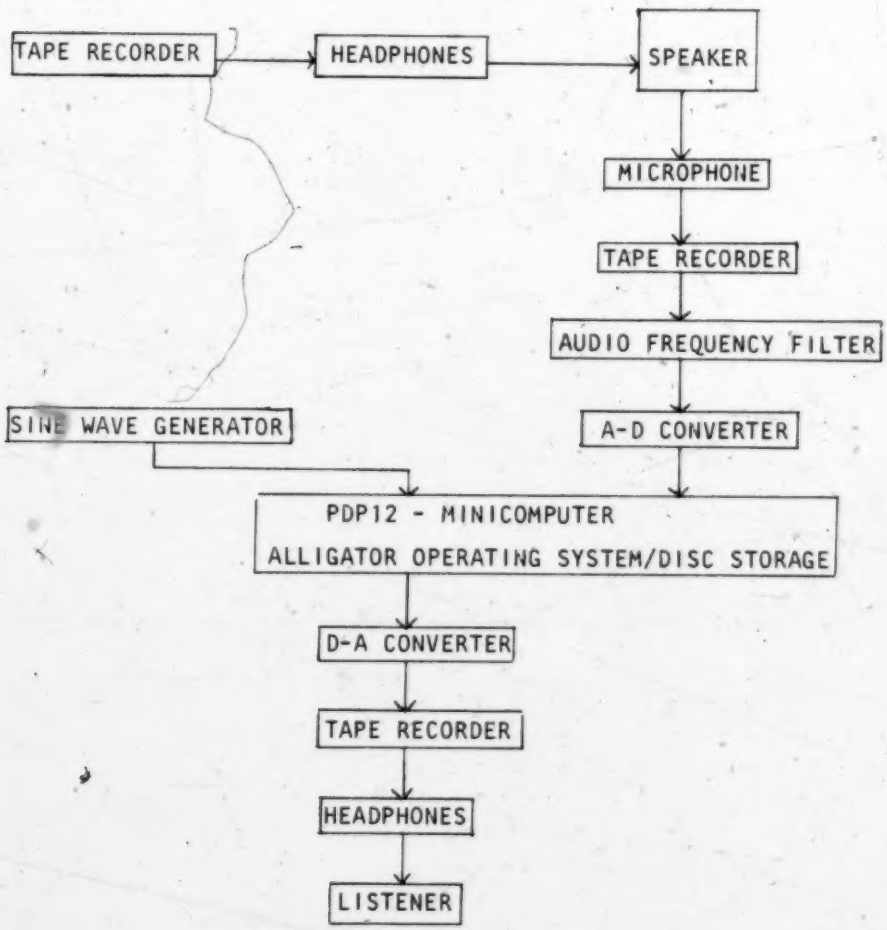
[1]Alligator is an operating system written in OS/8 PAL-12D and designed for use on a PDP-12 minicomputer. This system is designed for the manipulation and presentation of stimuli used in psychoacoustic experimentation (Stevenson and Stevens, 1978a,b).

vowels. Prior to recording each speaker was asked to read the list aloud. None of the speakers had any difficulty in following the instructions. Several recordings were made, discarding any tokens which either the speaker or the experimenter felt to be mispronunciations of the intended vowel. Prompting stimuli were presented with a SONY tape recorder over SONY headphones. (A wiring diagram is shown in Figure 1).

The stimuli were bandpass-filtered to eliminate frequencies below 68 Hz and above 6.8 KHz. Next they were sampled and digitized on the PDP-12 using the Alligator system. The stimuli were stored individually in 20 Alligator disc files: each file contained a set of isolated vowels or / pVp /'s from one speaker.

A presentation program was written in Alligator to generate a tape containing isolated vowels and / pVp / syllables from the first three speakers (two males and one female). The stimuli were passed through a desampling filter (bandpass 68 Hz to 6800 Hz) before being recorded onto audio tape. The set of isolated vowels and / pVp /'s from each speaker was randomized and presented in blocks of five, separated by a short tone, a sine wave of 1000 Hz. A 5-second inter-stimulus interval was used. Two replications of each speaker's vowels were recorded, for a total of 120 tokens. The vowels were presented in three sets of 40 vowels, one set for each speaker; 20 vowels in / pVp / frame followed by 20 isolated vowels. Within each condition the

Figure 1. block diagram of stimulus preparation and presentation



vowels were randomized, but speakers were "blocked", or segregated: all of the vowels from one speaker were presented before going on to the next speaker.

Procedure

Each listener was seated in a booth with a set of headphones, an attached microphone, and an answer sheet with 120 lines of the following response alternatives:

heed hid hayed head had hud hod hoed hood who'd

Subjects were instructed to listen to each sound, repeat what they heard into the microphone, then cross off the word on the response list which contained the same vowel. Before the test, the experimenter pronounced each of the ten vowels in both contexts, indicating the correct alternative from the list of response alternatives on the blackboard at the front of the room. Listeners were told that they would hear three different speakers throughout the experiment.

In this study, the order of presentation of the two context conditions (isolated vowels and / pVp / syllables) was not counterbalanced (although subsequent experiments, described below, control for this factor). Isolated vowels are unfamiliar stimuli, and it was felt that they might present special difficulties for phonetically naive listeners. Consequently each speaker's / pVp / syllables preceded the set of isolated vowels.

Tape recordings of each subject's spoken responses were made. Separate tapes and response sheets were collected from

each subject. Errors on the written responses were tabulated and recorded along with the incorrect response choices. For each vowel incorrectly identified in the written task, the listener's spoken response to the same stimulus was located on the tape. Each spoken vowel corresponding to a misidentified written vowel was re-recorded onto another tape. Three phonetically trained listeners, who were also longstanding residents of the Edmonton area, independently transcribed the set of vowels using a standard set of phonetic transcription symbols:

/ i, ɪ, e, ɛ, æ, ʌ, ɒ, o, ʊ, u, /

Majority responses were taken as the subject's intended response; that is, if 2 or 3 of the transcribers agreed on a given transcription it was taken to indicate the subject's spoken response.

In most cases, transcriber agreement was unanimous. In a few cases the listener either failed to give a spoken response, or majority agreement was not obtained. Such cases were treated as errors.

Error scores were recorded for all 120 vowels in the two conditions, written and spoken.

Due to time limitations, it was not possible to give listeners a practice session. Some subjects initially had difficulty with the instructions; others hesitated to begin the repetition task. Because of the large number of omissions near the beginning of the experiment, the first forty tokens (those of the first speaker) were treated as a

practice session, and only tokens 41 through 120 were analyzed.

Of the 26 listeners who participated in the study, eight were omitted, for the following reasons:

1. Four listeners failed to give spoken responses for the majority of the vowels;
2. Two of the listeners experienced a tape recorder malfunction;
3. Two of the listeners were not native speakers of English.

All subsequent analyses and discussion are based on the data for the last 80 tokens from the remaining 18 listeners. Error scores, averaged across listeners and speakers, are presented in Table 1. An error is considered to be any response other than the one intended by the speaker, including the absence of a response.

Results and Discussion

Comparing the two response conditions, written and spoken, it can be seen that very large differences emerge: a mean error rate of 16.18 percent for the written responses but only 4.72 percent for the spoken responses.

However, the large differences due to consonant context reported by Strange et. al. (1976) do not emerge in this study: / pVp / syllables averaged 15.14 percent errors in the written condition, compared to 17.22 for the isolated vowels. This rather small difference disappears in the

Table 1. Experiment 1: vowel identification errors

	i	ɪ	e	ɛ	æ	ʌ	ɒ	o	ʊ	u	Total
S1 C/W	6	0	0	0	0	18	3	2	16	9	54
S1 C/S	6	0	0	0	0	14	1	0	2	0	23
S1 V/W	0	1	0	2	1	22	10	0	15	3	54
S1 V/S	0	0	0	1	0	9	0	0	2	0	12
S2 C/W	1	0	1	0	0	12	11	0	19	11	55
S2 C/S	1	0	0	0	0	8	2	0	0	1	12
S2 V/W	3	2	0	0	0	28	11	1	20	5	70
S2 V/S	3	1	0	0	0	15	2	0	0	0	21

Legend:

S1: speaker 1
S2: speaker 2
C: vowels in /p_p/ context
V: vowels in isolation
W: written responses
S: spoken responses

spoken condition, where 4.86 percent errors are made for / pVp / syllables and 4.58 on the isolated vowels.

Since a large number of cells contained values of zero, the ANOVA model appeared to be inappropriate for these data. However, to test the hypothesis of orthographic interference it is unnecessary to analyze each vowel separately. Instead, the vowels were separated into two classes depending on their orthographic status: ambiguous vs. unambiguous. Table 2 shows the basis for this subcategorization. If a vowel had more than one common spelling form, and if more than one of these alternatives appeared on the response sheet, it was classed in the 'ambiguous' category. Otherwise, it was classified as 'unambiguous'. The former contains the vowels / i , ʌ , ɔ , ʊ , u , /; the latter contains the vowels /I, ɛ , e, æ, o, /

A factorial ANOVA was conducted on the data, which were pooled across speakers , repetitions and vowels, treating listeners as replications. Table 3 presents the results of this analysis. There are three main effects: orthographic confusability, response mode (written or spoken), and consonant context (vowels in isolation or / pVp / syllables).

Significant main effects were found for orthographic confusability and response mode ($p<.001$) and for the interaction of these two factors ($p<.001$), but not for consonant context. A simple main effects test on this interaction revealed significant improvement from the

Table 2. Examples of orthographically ambiguous words

Intended Vowel	Correct Keyword Response	Keyword Exhibiting Alternative Spelling	Examples of Alternative Spelling
i	heed	head	read, wheat
	hud	hood	flood, blood
		had	indefinite article 'a'
		hod	won, done
	hod	had	wad, call
U	hood	hud	put, full
u	who'd	hood	pool, food

Table 3. Experiment 1: ANOVA with simple effects test

SOURCE	df	Sum of Squares	Mean Square	F
A	1	10.5625	10.5625	0.34
B	1	1701.56	1701.56	54.78***
C	1	4795.56	4795.56	154.38***
AB	1	18.0525	18.0525	0.58
AC	1	00.0625	00.0625	0.00
BC	1	1387.56	1387.56	44.67**
ABC	1	10.5625	10.5625	0.34
within	8	246.500	31.0625	
Total	15	8170.42		
A at b1	1	3081.125	3081.125	99.199***
A at b2	1	8.0	8.0	.2576
B at a1	1	5671.125	5671.125	182.586***
B at a2	1	512.00	512.00	16.484**

*** p .001
** p .01

Legend:

A: response mode (written vs. spoken)
B: orthographic status (ambiguous vs. unambiguous)
C: consonant context (vowels presented in isolation or in /p_p/ context)

written to the spoken condition for orthographically ambiguous vowels, but not for the unambiguous set. As predicted by the hypothesis of orthographic interference, ambiguous vowels receive significantly higher error rates than unambiguous vowels. However, there is still a significant, though reduced difference between ambiguous and unambiguous vowels in the spoken condition. Examination of the confusion matrix (Table 4) suggests that a large proportion of those errors involve the vowel /ʌ/. Higher error rates for this vowel may reflect a genuine perceptual difficulty in addition to labeling difficulties. None of the remaining vowel pairs in the "ambiguous" class show any errors in the spoken condition.

One of the pairs displaying the potential for orthographic confusion, (/ɪ - ɛ/), does not occur as a confusion error under any condition. One possible explanation for this may be a left-to-right list-reading strategy; if listeners tend to select the first plausible candidate as the correct spelling for the vowel, they would select 'heed' rather than 'head' in response to the vowel /i/, because it appears earlier on the list.

These results indicate that orthographic interference may have a considerable effect on the outcome of vowel identification experiments. The increase in errors in the written condition is significantly greater for vowels which are orthographically ambiguous.

Further support comes from an examination of the

Table 4. Experiment 1: vowel confusion matrix.

Vowel Presented		Vowel Response: i C	i V	ɪ C	ɪ V	e C	e V	ɛ C	ɛ V	æ C	æ V	ʌ C	ʌ V	ɒ C	ɒ V	o C	o V	ʊ C	ʊ V	u C	u V	ø C	ø V
i	W	65	69			7	2								1								
	S	65	69			7	3																
ɪ	W			72	70				2														
	S			72	70				2														
e	W					71	72			1													
	S					72	72																
ɛ	W						1	71	70													1	1
	S							72	72														
æ	W								1	72	71												
	S									72	72												
ʌ	W							1		9	33	42	23	17	10			1				2	6
	S									2	7	50	48	13	8			1	2			6	7
ɒ	W						1	1		4	7	4	6	58	51	1	1	2				2	6
	S											1	1	69	70							2	1
o	W															70	70	1	1	1	1		
	S															72	72						
ʊ	W				1			1	5			18	15	4	1	2	1	37	36	1	1	9	12
	S															1		70	69		1	1	2
u	W											1		1		2	1	15	6	53	70		1
	S																			72	72		

Legend:

W: written responses
S: spoken responses
C: vowels in /p_p/ context
V: vowels in isolation
ø: no response or transcriber disagreement

confusion matrix (Table 4). A small number of confusions account for most of the errors.

Large differences between isolated vowels and / pVp / syllables are not obtained in this study. However, order of presentation was not counterbalanced. It is possible that practice with / pVp / syllables from a given speaker improved the identification rate for isolated vowels.

A second factor is the type of response required. Failure to obtain the results of Strange et. al. (1976) may be due to differences in the response frame used, since in their experiment listeners responded to / pVp / syllables and isolated vowels by selecting alternatives from a list of 'pVp' words: eg. peep, pip, pep, etc. For the CVC's this amounts to a simple spelling response; for the isolated vowels it involves a keyword response. In the written condition of the present study the response list consisted of 'hVd' words (ie. only keyword responses to both types of stimuli).

A second experiment was conducted to assess the role of these factors in determining consonant context differences.

Response frame and consonant context

In the preceding experiment it was shown that vowel misidentifications increase in number when more than one of the response alternatives contains an acceptable English spelling for the vowel presented. This effect emerges regardless of whether or not the stimulus vowel is presented

in isolation or in / pVp / frame.

B. Experiment 2: Keyword versus spelling responses

The present study was designed to assess the role of different response frames on the identification of vowels in isolation and / pVp / frame. In particular, it attempted to reconcile the findings of the first experiment with those of Strange et al. (1976) with regard to the effects of consonant context.

In this study a larger number of speakers was used (five males and five females). Listeners were divided into 4 groups. Half of the listeners (2 groups) heard the / pVp / syllables before the isolated vowels; the other half heard the isolated vowels first.

Half of the listeners within each order group were required to respond using answer frames which were 'pVp' words; the other half with answer frames which were 'hVd' words. Under the hypothesis that spelling responses are easier than keyword responses, it was predicted that isolated vowels would fare worse than CVC syllables when a spelling response was required for the latter, and a keyword response for the former; this difference should disappear when only keyword responses are used. An increase in errors from one response frame to the other can be attributed to

labeling difficulties rather than perceptual difficulties as such. Listeners using the / pVp / frame should obtain lower error rates when presented with / pVp / syllables (a spelling response) than when presented with vowels in isolation. Listeners responding with / hVd / words (a keyword response) should show an increase in errors in both contexts as a result of spelling or labeling biases. If a genuine perceptual advantage is afforded by the presence of consonant context, there should be a decrease in CVC errors regardless of the response frame. If the advantage is due to spelling interference or labeling difficulties, the difference should appear only for subjects responding with 'pVp' words.

Vowels were presented in the 'mixed' condition (i.e. with speakers randomized) in this study because it was anticipated that this condition would result in higher error rates (Strange et. al., 1976).

Listeners

Listeners were 18 students in an undergraduate course in practical phonetics. They had all received training in the use of phonetic symbols. Each was asked to complete a questionnaire concerning his/her language background (the same questionnaire was used in Experiment 1). Nine subjects were assigned to each group.

Stimulus materials

The data described in the previous experiment were again used in this study. An Alligator presentation program was used to generate two tapes containing 100 isolated vowels and 100 / pVp / syllables. Each set consisted of 10 tokens from each of ten speakers. Order of presentation was varied: one tape contained the / pVp / syllables followed by the isolated vowels; the second tape reversed this sequence. Vowels and speakers were randomized (the "mixed" speaker condition), with the constraint that no vowel or speaker was repeated on adjacent trials. The stimuli were presented in blocks of five, with a five second interstimulus interval. An additional pause of three seconds was inserted at the end of each block to help subjects keep their place on the list.

Apparatus

The experimental apparatus for the preparation of stimuli was identical to that described in the first experiment. The following instruments were used for stimulus presentation:

1. Amplifier: SONY TA-1066.
2. Tape recorder: TEAC A-7030.
3. Loudspeaker: HECO Sound Master 15.
4. Audio frequency filter: Rockland 1524-01.

Procedure

The experiment was conducted in a quiet seminar room. The stimuli were presented over a HECO loudspeaker using a

TEAC recorder and a SONY amplifier. Listeners were seated at tables encircling and approximately equidistant from the sound source. The amplitude was adjusted so that the stimuli were clearly audible in all parts of the room.

Two kinds of response booklets were prepared. One contained 200 lines of 'pVp' words:

peep pip pape pep pap pup pop pope puup poop

The other contained 200 lines of 'hVd' words:

heed hid hayed head had hud hod hoed hood who'd

Half of the subjects within each of the two groups used the first response frame; the other half used the second. Subjects were instructed to cross off the word which contained the vowel they heard. They were asked not to go back or omit any responses, picking the most likely answer in case of doubt. The experimenter repeated the set of isolated vowels and CVC syllables twice, pointing to the correct written response (for each response frame) on the blackboard.

Listeners were informed that they would hear 100 CVC syllables and 100 vowels in isolation from a number of speakers in random order. A brief rest pause was given at the end of each set. After the first group of subjects completed the task, the second group was ushered in and the entire procedure was repeated.

Results and Discussion

None of the listeners had any difficulty with the task. The data from 4 listeners were omitted for reasons of dialect and/or language background.

The data were pooled across the two presentation orders. Since very few or no errors were made for a number of vowels, it was necessary to pool across the vowels. Because the present experiment was not directly concerned with differences between the speakers, the data were collapsed across this factor as well. A summary table is presented in Table 5.

An analysis of variance was conducted using these data on the factors Context, Response frame, and Listeners-within-frames (see Table 6). A partially repeated measures design was used: listeners were nested within response condition (/ pVp / or / hVd /) but crossed with respect to context (/ pVp / vs. isolated vowels). Orthogonal comparisons were set up to test the following hypotheses:

1. Are error scores significantly lower when the response required is a spelling response rather than a keyword response? Ie. are there fewer errors in the case where listeners are responding with 'pVp' words to / pVp / stimuli?
2. Isolated vowels necessarily require keyword responses. Is there still an advantage due to context (ie. are more errors made on isolated vowels in either response condition) when responses to / pVp / syllables are also of the "keyword" type?

Table 5. Experiment 2: vowel identification errors.

	hVd		pVp	
Listener	CVC	ISO	CVC	ISO
1	13	10	2	7
2	6	7	5	2
3	12	13	1	8
4	4	6	8	13
5	7	12	6	15
6	3	5	2	9
7	2	5	3	8
8	7	18	4	13
9	23	13	5	11
10	20	12	4	3
11	2	24	3	13
12	2	11	3	13
13	6	3	4	11
14	3	10	4	2

Legend:

hVd: listeners responding with hVd frame
pVp: listeners responding with pVp frame
CVC: vowels presented in /p_p/ context
ISO: vowels in isolation

Table 6. Experiment 2: Partially repeated measures ANOVA with planned comparisons

SOURCE	df	Sum of Squares	Mean Square	F
C1	1	301.339	301.339	14.216**
C2	1	15.750	15.750	0.74
C3	1	38.679	38.679	1.83
L(F)	26	734.605	28.254	
CL(F)	26	553.745	21.298	
Total	55	1644.119		

**p .01

Legend:

C1 to C3: see text
L: listeners
F: frame (listeners responding with pVp vs. hVd frame)
C: context (vowels presented in isolation or in /p_p/ context)

3. Are isolated vowels more frequently misidentified than / pVp / syllables when 'hVd' responses are given?

A significant F ratio was obtained for the first contrast described above ($F = 14.31$; d.f. = 1, 26; $p<.01$). Neither of the other contrasts was significant.

These results show that listeners perform best when they are required to give spelling responses. A significant increase in errors is found for keyword responses to both isolated vowels and / pVp / syllables.

Vowels in isolation fare slightly worse than / pVp / syllables when a keyword response is required. This difference is not statistically significant.

In spite of the fact that more speakers were employed in this experiment, somewhat lower error rates were obtained compared with the results reported in the previous chapter (an average of 9.21 percent errors across all conditions involving keyword responses, compared with 16.18 per cent in the written condition of Experiment 1.) Phonetic training may reduce the amount of interference from orthography, or in some other way facilitate the identification of a word from a response list. It is of considerable interest therefore to find that labeling difficulties are present even when subjects have been introduced to another transcription system and are made aware of the lack of one-to-one correspondence between vowels and their othographic representations.

Examination of confusion matrices (see Table 7)

Table 7. Experiment 2: vowel confusion matrix

Vowel Response

Vowel Presented		i		ɪ		e		ɛ		æ		ʌ		ɑ		o		ʊ		u		ø	
		C	V	C	V	C	V	C	V	C	V	C	V	C	V	C	V	C	V	C	V	C	V
i	P	139	138	1	2																		
	H	135	136					5	4														
ɪ	P		1	137	132			1	7									2					
	H		1	139	130		1	1	5										3				
e	P	14	2			126	138																
	H	13	2	1		125	136	1	1										1				
ɛ	P				4		1	136	133	4	1												1
	H			4	8	2	2	132	125	2	3		1						1				
æ	P						1	1	3	139	136												
	H					2	1		1	135	135	2	1	1									2
ʌ	P								1	7	39	122	88	10	11			1	1				
	H				2	3	1	1	3	11	28	102	80	19	24		1	3	1			1	
ɑ	P										6	2	22	138	110				1		1		
	H					4	1			1	2	7	15	127	121			1	1				
o	P												1			139	139			1			
	H												1		3	139	136			1			
ʊ	P			1					1		1	3	11			1		130	126	3	1	2	
	H			3	2			1	1			6	10	4	4			121	120	3	2	2	1
u	P																		8	140	132		
	H																	5	8	135	132		

Legend:

P: listeners responding with pVp frame
H: listeners responding with hVd frame
C: vowels in /p_p/ context
V: vowels in isolation
ø: no response

indicates that many of the same errors appear in both experiments 1 and 2. An analysis of vowel confusions in terms of spelling ambiguity is complicated by the fact that different spelling confusions are predicted for / pVp / and / hVd / response words. Confusions between / ʌ - æ /, / i - ɛ /, / u - ʊ /, and / ʌ - ʊ / (all of which display the potential for spelling interference : see Table 2) show substantially higher error rates when keyword resonses are given than when spelling responses are required. Other vowel pairs which show a similar reduction in errors in the spelling condition are / ɛ - ɪ / and / ɒ - ʌ /.

Some vowel pairs occur as substitution errors more frequently in the isolated vowel condition than in the / pVp / condition: these include / ʌ - æ / and / ɪ - ɛ /. One pair, / e - i /, occurs frequently as a substitution error for / pVp / syllables but not for isolated vowels. These errors may reflect genuine perceptual ambiguities. However, conclusions with respect to the perceptual confusability of vowel pairs are premature as subjects appear to be having difficulty assigning the correct labels to vowels.

Since it is difficult (if not impossible) to separate orthographic errors from perceptual errors, an attempt is made in subsequent experiments to eliminate spelling interference entirely.

III. PERCEPTUAL STUDIES ON THE ROLE OF CONTEXT

In previous studies it has been demonstrated that certain kinds of response requirements may bias the outcome of vowel identification experiments. While a more detailed investigation of such factors as orthographic interference may be of value, the present study is concerned only with simplifying the task for the listener in order to circumvent labeling problems.

There are several possible solutions to the labeling problem. One solution involves having listeners repeat the vowels rather than transcribe them. Transcription is then carried out by a panel of trained listeners who are speakers of the same dialect. This procedure carries the implicit assumption that vowel categories are aligned between listeners and transcribers. One consideration is that the recording and transcription procedure involves a great deal of time and effort.

A second possibility is the use of a standard transcription system which is not directly linked to the orthography. It would be necessary to use listeners who are fully experienced in the use of the symbols: assignment of a symbol should be an automatic process.

Another method involves the use of orthographic responses in conjunction with a moderate amount of practice

to overcome labeling difficulties. This alternative may be undesirable if perceptual learning takes place. For example, a listener may become familiar with the individual tokens or voices.

The training session should therefore consist of tokens not actually used in the experiment itself. In the present study the second and third methods were both adopted. Listeners were experienced in the use of phonetic symbols and in spelling/sound correspondences. Both orthographic symbols and phonetic symbols were used as responses to determine which procedure yields better results. It is assumed that any difference between the two types of responses can be attributed to labeling difficulties and has no bearing on the actual identification process.

A third experiment was designed using the latter two methods to evaluate the role of consonant context on vowel identification.

A. Experiment 3: consonant context and vowel identification

This experiment compared the identification of vowels in isolation and / pVp / frame for two response types: spelling forms and phonetic symbols. The spelling frames were the same as those used in Experiment 2, thus allowing a further comparison of keyword and spelling responses. In

order to determine whether labeling difficulties and orthographic interference disappear with practice, the same set of listeners was used, but they were divided into different groups. In addition to crossing off a word from the response list, each listener was required to first transcribe the vowel using a standard transcription system (the IPA, or International Phonetic Alphabet) with which they were familiar. It was hoped that the IPA system would be less likely to generate spelling or labeling errors, and hence provide a fairly "neutral" method for evaluating the effects of consonant context. It was anticipated that fewer errors would occur using a transcription system which is not directly linked to the English spelling system.

Listeners

A total of 28 listeners took part in the study. All but 2 of the listeners who took part in the present study also completed Experiment 2. They were all familiar with the transcription system and presumed to be adept in the transcription of vowels using IPA symbols. Seven subjects were assigned to each of the four groups corresponding to the two orders of stimulus presentation and two response modes, as they were in the second experiment.

Stimulus materials

The set of 100 isolated vowels and 100 / pVp / syllables were again used in this study. The stimuli were

prepared exactly as in the second experiment.

Apparatus

A relatively noise-free seminar room was used to conduct the experiment. A TEAC tape recorder was used, in conjunction with a SONY amplifier and a HECO loudspeaker. Subjects were seated at approximately equal distances from the loudspeaker.

Procedure

The listeners were given a questionnaire to complete regarding their language background, as in the previous experiments. The response booklets were identical to those used in the second experiment: half of the subjects were given lists of 'pVp' words, the other half 'hVd' words.

Instructions were somewhat different: listeners were told that they would hear vowels in isolation and in / pVp / frame. They were instructed to listen to the vowel, write down the IPA symbol corresponding to the vowel and then mark off the word from the list of ten alternatives. Subjects were instructed not to skip any answers, selecting the most likely alternatives in case of doubt. The response alternatives were as follows:

IPA: / i, ɪ, eɪ, ɛ, æ, ʌ, ɒ, oʊ, ʊ, u /

'pVp': 'peep pip pape pep pap pup pop pope puup poop'

'hVd': 'heed hid hayed head had hud hod hoed hood who'd'

They were expressly told not to cross off the word until

they had assigned the transcription label. This was done to avoid orthographic interference from the spelling task. The IPA symbols were written on the blackboard at the front of the room, and were pronounced by the experimenter.

The experiment was conducted in two sessions, corresponding to each order of presentation. Fourteen subjects were assigned to each order group. Seven subjects were assigned to each response frame ('hVd' or 'pVp' words) in both order groups. Groups were matched as closely as possible for language background and city or region of origin, based on information in previous questionnares.

Three of the listeners encountered problems with the task. Evidently they were not completely fluent in the use of IPA symbols since they failed to provide any IPA symbol for a large number of items. For this reason their data were omitted from subsequent analyses. Three additional listeners were omitted for reasons of dialect and/or language background.

Results and Discussion

Separate analyses were conducted for spelling responses and IPA responses.

Spelling and keyword responses

The data were collapsed across vowels and talkers as well as the two presentation orders. Table 8 shows the mean error rates for each stimulus and response type. As

Table 8. Experiment 3: vowel identification errors.

	hVd		pVp	
Listener	CVC	ISO	CVC	ISO
1	7	5	11	20
2	15	5	2	8
3	5	7	7	24
4	5	4	4	17
5	5	3	4	10
6	7	8	3	2
7	2	5	1	4
8	3	9	4	3
9	6	6	2	2
10	3	5	2	7
11	1	11	2	7

Legend:

hVd: listeners responding with hVd frame
pVp: listeners responding with pVp frame
CVC: vowels presented in /p_p/ context
ISO: vowels in isolation

predicted, spelling responses (responding with 'pVp' words to / pVp / stimuli) result in the lowest error rates, as found in the previous experiment. However, errors in the 'hVd' response condition are lower for both types of stimuli.

Since each listener heard both isolated vowels and / pVp / syllables but used only one response frame, a partially repeated measures ANOVA was used. The results of this analysis can be seen in Table 9. As in Experiment two, a priori orthogonal contrasts were set up. The first contrast compares the first quadrant (spelling responses) against all of the others (keyword responses). This contrast is no longer statistically significant, even though there is still a trend in the predicted direction. Neither of the other contrasts was significant. CVC syllables do not fare better than isolated vowels when only keyword responses are used.

Examining the confusion matrix in Table 10, it appears that some of the same confusion errors emerge in this study as in the previous one: / e - i / confusions are common for / pVp / syllables; / ʌ - æ /, / ʌ - ɒ / and / ɒ - ʌ / errors are frequent for isolated vowels. Two of the three keyword response conditions show a considerable increase in the number of omissions (ie. failure to give a response) over the spelling response condition. This is consistent with the hypothesis that the spelling task is easier to perform. However, the factor of orthographic interference may be

Table 9. Experiment 3: Partially repeated measures ANOVA with planned comparisons

SOURCE	df	Sum of Squares	Mean Square	F
C1	1	83.5227	83.5227	3.5811
C2	1	58.9091	58.9091	2.5258
C3	1	44.1818	44.1818	1.8943
L(F)	20	390.090	19.5045	
CL(F)	20	466.451	23.3226	
Total	43	1043.154		

Legend:

C1 to C3: see text
L: listeners
F: frame (listeners responding with pVp vs. hVd frame)
C: context (vowels presented in isolation or in /p_p/ context)

Table 10. Experiment 3: vowel confusion matrix

		i		ɪ		e		ɛ		æ		ʌ		ɒ		o		ʊ		u		ø	
		C	V	C	V	C	V	C	V	C	V	C	V	C	V	C	V	C	V	C	V	C	V
i	P	109	109						1													1	
	H	109	106		2				2													1	
ɪ	P		2	110	101				5										1				1
	H			108	109				1													2	
e	P	70				100	108		1														1
	H	18				91	110															1	
ɛ	P			1	2			109	104		3												1
	H			2	7			107	98		4								1			1	
æ	P							2	1	104	105	2	2	1				1					2
	H							2	3	101	106	3		2	1							2	
ʌ	P									1	21	101	72	7	13			1					4
	H									3	13	98	84	7	12							2	1
ɒ	P										1	3	6	107	95				5				3
	H				1							2	13	107	95		1					1	
o	P														1	109	107		1	1			1
	H															110	109						1
ʊ	P				2							5	8			1		102	95	2	2		3
	H				1							3	1			1		102	108	4			
u	P																	3	9	107	100		1
	H																	2	3	108	107		

Legend:

P: listeners responding with pVp frame
H: listeners responding with hVd frame
C: vowels in /p_p/ context
V: vowels in isolation
Ø: no response

reduced in the present study: / ʊ - u /, / ʊ - ʌ /, / ɒ - æ / and / u - ʊ / errors do not show an increase in the keyword condition, as predicted earlier.

The average error rate for 'hVd' responses is 5.77 percent, compared with 9.25 percent for the comparable condition in Experiment two. With more practice, listeners may be able to overcome labeling difficulties and perform the keyword task as well as they perform the spelling task. The results for the group responding with 'pVp' words to isolated vowels suggest that this is the hardest task: there was no improvement from the previous experiment in this condition (9.14 percent as compared with 9.46). The 'pVp' frame may be more difficult as a keyword response. Alternatively, the lack of improvement may be due to differences between subject groups.

IPA responses

All listeners performed the IPA task in addition to responding with 'hVd or 'pVp' words. Since it is impossible to rule out a transference effect from the other task, the IPA data were not collapsed over the response groups. Despite instructions to the contrary, some subjects may have crossed off their response from the word list before writing down the IPA symbol.

Error scores were tabulated, collapsing across orders, vowels and talkers (see Table 11). A partially repeated measures ANOVA was conducted on the factors stimulus type

Table 11. Experiment 3: vowel identification errors (IPA responses)

	hVd		pVp	
Listener	CVC	ISO	CVC	ISO
1	7	5	11	17
2	13	5	1	5
3	6	5	12	17
4	4	4	4	17
5	5	3	3	10
6	8	10	14	7
7	1	5	1	4
8	3	7	3	3
9	6	5	2	2
10	3	5	2	5
11	1	13	2	7

Legend:

hVd: IPA responses by listeners using hVd frame
pVp: IPA responses by listeners using pVp frame
CVC: vowels presented in /p_p/ context
ISO: vowels in isolation

(/ pVp / syllables or isolated vowels) and response group (listeners responding with either 'hVd' or 'pVp' in the spelling task). Since subjects in both response groups used IPA symbols, it is assumed that any difference between these two groups is due to interference from the first task. Table 12 shows the results of this analysis. A significant F-ratio was obtained for stimulus type ($p<.05$), but not for response group or the interaction of the two factors. Isolated vowels obtained slightly higher error rates than CVC syllables (7.32 as compared with 5.09 percent).

Overall error rates are higher for IPA responses (6.21 percent errors) than for spelling responses (3.82 percent), but very similar to keyword responses in Experiment 3 (5.77 percent).

The confusion matrix appears in Table 13. Isolated vowel pairs which account for greater than 5 percent of all errors include / ʌ - æ /, / ʌ - ɒ /, / ɒ - ʌ /, / ʊ - ʌ /, / ɛ - ɪ /, / ɛ - æ /, and / ɑ - ʊ /. Errors for / pVp / stimuli which account for more than 5 percent of all errors are: / ʌ - ɒ /, / ɒ - ʌ /, / ʊ - ʌ / and / ʊ - ɑ /. The most striking differences between the two conditions involve / e - i / confusions (none in the isolated vowel condition), and the increase in errors for the vowel / ʌ / in the isolated vowel condition (in which / ʌ - æ / and / ʌ - ɒ / confusions predominate).

In conclusion, it would appear that some differences exist between / pVp / syllables and isolated vowels.

Table 12. Experiment 3: Partially repeated measures ANOVA on IPA data

SOURCE	df	Sum of Squares	Mean Square	F
F	1	14.2045	14.2045	0.54*
C	1	54.5682	54.5682	4.39
L(F)	20	530.453	26.5226	
FC	1	19.1136	19.1136	1.54
CL(F)	20	248.813	12.4406	
Total	43	867.152		

*p .01

Legend:

F: frame (listeners responding with pVp vs. hVd frame)
C: context (vowels presented in isolation or in /p_p/ context)
L: listeners

Table 13. Experiment 3: vowel confusion matrix

Vowel Response

Vowel Presented		i C	i V	ɪ C	ɪ V	e C	e V	ɛ C	ɛ V	æ C	æ V	ʌ C	ʌ V	ɔ C	ɔ V	o C	o V	ʊ C	ʊ V	u C	u V	Ø C	Ø V
i	H	110	109		1																		
	P	110	109						1														
ɪ	H			110	109				1														
	P		2	110	100				6										1				1
e	H	17				93	110																
	P	10				100	108		1														1
ɛ	H			2	8			103	97		4								1				
	P			1	1			108	104		4											1	1
æ	H							1	2	104	104	3		2	4								
	P							1		105	108	2						2					
ʌ	H									3	13	96	85	6	11	1		1				3	1
	P										19	101	74	7	13			1				1	4
ɒ	H											4	13	105	95				2			1	
	P										1	3	2	105	102				4	1		1	1
o	H															110	110						
	P														1	109	108	1					1
ʊ	H				1							4	1	1	1	2		99	107	4			
	P				2							5	11					96	90	4	4	4	3
u	H																	2	3	108	107		
	P											1					1	3	5	101	103	5	1

Legend:

H: IPA responses by listeners using hVd frame
P: IPA responses by listeners using pVp frame
C: vowels presented in /p_p/ context
V: vowels in isolation
Ø: no response

However, these differences are rather minimal when compared with the results of Strange et. al., and suggest that the importance of consonant context may have been exaggerated in the past. The present results do not support any theory which maintains that consonant context is essential in the perception of vowels. In particular, they do not support the view that consonant transitions are essential for the specification of the vowel. Information present in the vowel is not inseparable from that of the consonant. Rather, the results suggest that context-free randomized vowels are fully specified, well-recognized stimuli, in spite of the fact that vowels do not ordinarily occur in isolation, and the lax vowels never appear in word-final position. Since the actual tokens differ from one context condition to the other, the possibility of errors in production cannot be ruled out. In the case of speaker context, however, errors in production are ruled out since the actual tokens are the same; only the order of presentation is different.

These results also indicate that the use of IPA transcription by trained listeners may offer an less biased method for evaluating the effects of context. The use of keyword frames is not recommended unless the listeners have had considerable experience with the task.

The error rate for isolated vowels (7.32 percent) is extremely low, considering that a vowel in connected speech is provided with consonant and speaker context as well as semantic, syntactic and phonological information. The next

question which arises is whether any additional improvement is afforded by the presence of speaker context. A further experiment was conducted, comparing isolated vowels in the mixed condition and in the blocked condition.

B. Experiment 4: Speaker context and vowel identification

This experiment set out to determine whether speaker information (as provided by "context" vowels) can aid the vowel identification process. Previous experiments have indicated that isolated vowels are well recognized even in the absence of consonantal context. The present study was conducted to find out what proportion of the remaining errors can be eliminated by including speaker information. It is assumed that listeners have access to this information when all of a given speaker's vowels are presented in a block.

Naturally, the first vowel in each block will involve a shift in speaker identity. No assumptions are made concerning the nature and number of vowel tokens necessary for the listener to extract speaker information. Under most of the theories described above, exposure to several of a speaker's vowels should facillitate the identification of subsequent vowels, resulting in fewer misidentifications.

The IPA transcription system was used, to avoid the

possibility of orthographic interference. There is a fairly close concordance between the number of errors obtained in the first experiment (using spoken responses) and the third experiment (using IPA responses by phonetically trained listeners). In the first experiment listeners made fewer errors on isolated vowels. This experiment involved a smaller number of speakers, hence reduced inter-speaker variation; moreover, the vowels were presented in the blocked condition. The use of phonetically trained subjects seems justifiable on the basis of this result: their behaviour is not substantially different from that of untrained subjects.

Listeners

Subjects were 36 undergraduate students enrolled in a practical phonetics course. All of the listeners had taken part in one or both of the previous experiments. All were familiar with the use of IPA symbols.

Stimulus materials

The same set of 100 isolated vowels described in previous experiments served as stimuli. An Alligator program was written to prepare the stimulus materials, which were recorded onto two audio tapes. Randomization lists were generated using the Alligator with the following specifications:

1. Blocked condition: vowels from each speaker were

randomized, and presented sequentially (each speaker's vowels were presented in a block of 10).

2. Mixed condition: vowels were randomized with respect to vowel identity and speaker identity. The same vowel randomization was used in both conditions, to avoid possible sequence effects.

The only difference between the two conditions lies in the identity of the speaker. In the mixed condition it was constantly changing, while in the blocked condition it was fixed (speaker identity changed only at the end of each block). Two additional constraints were imposed: each of the vowels appeared once in every block of ten; and no pair of adjacent vowels were the same.

Each tape contained isolated vowels in both the blocked and the mixed condition. The first tape presented the vowels in the blocked condition followed by vowels in the mixed condition. The second tape reversed this order.

Apparatus

The experimental apparatus was similar to that used in previous studies, and is described above. Preparation of the experimental stimuli was undertaken with the PDP-12 minicomputer, using the Alligator system. The stimuli were passed through a desampling filter (68 to 6800 Hz) using the Rockland 1524-01 and subsequently recorded onto audio tape using a TEAC tape recorder.

Procedure

18 subjects were assigned to each order condition: mixed/blocked and blocked/mixed. Subjects in the two conditions were matched as closely as possible with respect to geographic area of origin and dialect.

The tests were conducted in two sessions in a relatively noise-free speech laboratory. A TEAC tape recorder was used with a SONY amplifier and a HECO loudspeaker. Listeners were seated around a table at approximately equal distances from the loudspeaker. The stimuli were clearly audible in all parts of the room. Each listener received an answer sheet containing numbers from 1 to 100. Listeners were told that they would hear 100 vowels from 10 different speakers. They were asked to transcribe each vowel using IPA symbols . The vowel symbols were written on the blackboard and pronounced by the experimenter. None of the listeners had any difficulty transcribing the vowels.

Results and Discussion

One subject's data were eliminated for reasons of language background. This individual was not a native English speaker and made a considerable number of errors.

The data were collapsed over the presentation orders, for a total of 35 observations in each context condition (see Table 14). A small improvement was found, in the direction predicted (5.43 percent in the mixed condition,

Table 14. Experiment 4: vowel identification errors

Listener	M	B
1	2	1
2	8	5
3	3	3
4	1	0
5	5	2
6	6	5
7	4	4
8	16	5
9	3	0
10	7	1
11	5	7
12	2	2
13	10	7
14	2	3
15	11	4
16	3	1
17	15	6
18	14	9

Listener	M	B
19	8	11
20	3	1
21	1	8
22	4	7
23	6	3
24	3	4
25	9	10
26	3	9
27	5	4
28	4	1
29	3	2
30	6	8
31	4	1
32	2	0
33	3	1
34	3	3
35	6	5

Legend:

M: Mixed speaker condition
B: Blocked speaker condition

4.09 in the blocked condition). A t-test for correlated means was conducted. A significant improvement was found for the blocked condition ($t(34)=2.016, p<.05$).

The confusion matrix (see Table 15) indicates that most errors involve substitutions of / ɪ - ɛ /, / ɛ - æ /, / ʌ - ɑ /, / ʌ - ɔ /, / ɔ - ʌ /, and / ʊ - ʌ /. Each of these pairs occurred as common errors in the previous experiments as well. All of these except / ʊ - ʌ / and / ʌ - ɔ / show some improvement in the blocked condition.

The vowels / i /, / e / and / o / are rarely misidentified, together contributing less than 3 percent to the overall error rate in either condition. The vowel / ʌ / showed the highest proportion of errors, contributing over 40 percent of the total number of errors in either condition.

It is significant that very few errors are made in identifying context-free, speaker-randomized vowels. This indicates that intrinsic properties of the vowels contain sufficient information to enable listeners to extract vowel quality. Theories which maintain that vowels are specified by their context are not supported.

Under a formant normalization hypothesis, context-free vowels are expected to be highly confusable, given the nature and extent of the overlap in their formant frequencies. The results obtained in the present study are not consistent with this prediction: very few errors are obtained for speaker-randomized isolated vowels. Context

Table 15. Experiment 4: vowel confusion matrix

Vowel Presented		Vowel Response: i	ɪ	e	ɛ	æ	ʌ	ɒ	o	ʊ	u	ø
i	M	179		1								
	B	179		1								
ɪ	M	1	161		14	1	2			1		
	B	1	170		9							
e	M			180								
	B	1		179								
ɛ	M		5	1	152	19	1	1				1
	B		2	1	166	11						
æ	M				3	177						
	B				1	178						1
ʌ	M				1	43	98	30		6		2
	B					29	118	30		1	1	1
ɒ	M					3	23	153		1		
	B						17	160		2	1	
o	M								180			
	B							2	178			
ʊ	M		4		3		12	2		157	2	
	B				9		12	1		153	2	3
u	M		2				1			4	173	
	B					1				3	176	

Legend:
M: Mixed speaker condition
B: Blocked speaker condition
ø: no response

appears to be less critical than other intrinsic parameters of the vowels. However, there is a second possibility which cannot be ruled out; the data used in these experiments may not exhibit the variations in formant frequencies that are indicated in other data samples. This issue will be dealt with in Chapter 4.

How is it that listeners can extract information from context-free vowels if these vowels exhibit large variations in their acoustic properties? One possibility is that listeners are also extracting information from other sources in the signal. Dynamic characteristics such as duration and diphthongization may provide information which helps to disambiguate vowels with overlapping formant frequencies. These parameters are realized in traditional phonetic descriptions as length and (non-phonemic) diphthongal offglides.

The English vowels are frequently grouped into two classes, tense and lax. Vowel pairs such as / i - ɪ /, / e - ɛ / and / o - ʊ / are adjacent in the vowel diagram but contrast in length and diphthongization (Joos, 1948).

Tense vowels are, on the average, longer than the corresponding lax vowels (Peterson and Lehiste, 1960). Diphthongal offglides for the tense vowels / i , e , o , u / have been described as tending toward the extremes of the vowel diagram (toward / i / for the front vowels, / u / for the back vowels). In many North American dialects of English another type of diphthongization has been observed for the

lax vowels /ɪ, ɛ, ʊ, æ, /. These vowels may show offglides in the direction of schwa / ə / (Joos, 1948). Such vowels ordinarily occur in preconsonantal position only; when spoken in isolation, schwa offglides may be more pronounced. In other syllabic contexts (for example, before / g /) and in certain prosodic contexts lax vowels may show diphthongal offglides as well (Avis, 1972).

While the acoustic analysis of diphthongization has not received much attention there is some indication that this property may be represented in terms of changing spectral characteristics. The gliding movement is characterized by changes in the formant frequencies of the vowel from its onset to its termination. Peterson and Coxe (1953) found that formant changes for / e / and / o / were most pronounced in stressed syllables, particularly open syllables, and in citation form. They note variability in the onset and termination positions of the formants, although the direction of change was constant.

Gay's (1977) study of the English phonemic diphthongs / ɔ̯ɪ , ɑ̯ɪ , ɑ̯ʊ / indicates that rate of second formant frequency change may be an important cue for distinguishing vowel pairs such as / ɑ - ɑ̯ɪ /. Using synthetic speech, he found that the preferred values for the onset and terminal positions of these vowels varied as a function of vowel duration, but that they could be characterized by an invariant rate of change in F2.

The perceptual role of vowel duration was investigated

by Ainsworth (1972) and Bennett (1968). Their findings show that duration may be an important factor for the discrimination of synthetic vowels which have similar formant frequencies (ie. neighbouring vowels in the F1-F2 space).

Lehiste and Meltzer (1973) synthesized a set of ten vowels on the basis of formant measurements from the vowels of an adult male, an adult female and a ten year old child. Fundamental frequency measurements were taken from the Peterson and Barney (1953) averages for men, women and children. The set of naturally spoken vowels were identified at much higher rates of identification (an improvement of approximately 25 percent) than the synthesized vowels. While it is not explicitly stated that dynamic characteristics were omitted from the synthetic tokens, it appears that spoken vowels contain additional sources of information not accounted for by the formant frequencies.

Context-free isolated vowels are not frequently misidentified, contrary to predictions based on earlier findings of extensive overlap between formant frequencies of different vowels. If vowels contain additional sources of information in the form of dynamic characteristics, these properties may help to separate vowels with non-unique formant patterns.

By eliminating such characteristics, it should be possible to induce a greater number of confusion errors. If so, it should also be possible to determine whether

listeners make use of the relational characteristics inherent in speaker context. Since listeners appear to utilize such information in synthetic speech, and since there is relational information potentially available in natural speech, it might seem logical that they would rely on such sources when additional information is not available. That is, if diphthongization and duration differences are neutralized, identification errors are expected to increase. If relational information is extracted from a set of vowels from the same speaker, fewer errors are expected in the presence of speaker context (the blocked speaker condition).

IV. ACOUSTIC AND PERCEPTUAL STUDIES OF GATED VOWELS

Previous studies have shown that isolated vowels are well recognized even in the absence of context. Listeners are able to identify such vowels in spite of formant overlap, which suggests that additional cues are involved. Information may be present in the form of dynamic characteristics like duration and diphthongization. By removing such characteristics from the signal, an increase in confusion errors should result for vowels which are spectrally similar.

There is some experimental evidence that error rates go up when vowels are artificially shortened by removing portions of the waveform. Tiffany (1953) spliced a series of sections from the central portions of 12 sustained vowels. These sections were 80, 200, 500 and 8000 msec in duration. Onset and termination portions were removed by means of tape splicing. These vowel sections were compared with isolated vowels and vowels in / tVp / frame with speaker-controlled durations of 200 msec. All vowels from the three conditions were randomized for each of the four speakers (ie. speakers were blocked). Eighteen trained listeners transcribed the vowels using phonetic symbols. Each vowel was presented twice.

Tiffany found that the vowels / e / and / ɑ / were more likely to be incorrectly recognized at shorter durations, while / ɪ / and / ʊ / were more often misidentified at longer durations. This is consistent with Peterson and Lehiste's (1960) observation that / ɪ / and / ʊ / are "intrinsically short" vowels. In general, "short" vowels tended to occur more frequently as incorrect responses to stimuli of short duration. / ɑ - ʌ / confusions are more common at shorter durations, while / ɑ - ɔ / confusions are increased at longer durations. The vowels / i /, / ɜ /, / æ /, and / u / were well recognized under all conditions. The vowels most affected by the splicing were / e / and / o /, which are most commonly diphthongized in connected speech. Tiffany suggests that diphthongization and duration may provide information which is important for the recognition of these vowels.

Tiffany's study did not examine the effects of speaker context. Considerably more confusions might be expected in a speaker-randomized condition, particularly for spectrally similar vowel pairs. The experiment described here was conducted to test this hypothesis and to determine whether errors are related in a straightforward way to formant overlap when dynamic information is eliminated from the signal.

Joos (1948) claims, on the basis of tape-splicing experiments, that diphthongization characteristics disappear at durations under 80 msec. Tiffany found an increase in

errors for vowel sections of this length. In the present study this duration was chosen as well.

Visual inspection of a number of sonagrams of the vowels used in this study indicated that the steady state portion generally occurred within the first 200 msec. It was decided to take sections following the first 100 msec. of each vowel.

A. Experiment 5: mixed and blocked speaker context and the identification of gated vowels

Listeners

Eight phonetically trained listeners completed the task. All but three of the listeners were native to Western Canada. Two of the remaining subjects came from other parts of Canada (Ontario and Newfoundland). One subject came from the eastern United States. All were familiar with the Edmonton area dialect. In addition, each had participated in the listening test of Experiment 4. The scores obtained on this test were comparable with the results of other participants in that study.

Stimulus materials

The set of 100 isolated vowels described in previous experiments was again used in this study. Each token was

adjusted by means of a window function which performed the following operations:

1. The initial 100 msec portion was effectively eliminated, multiplying each point by zero.
2. The second 10 msec portion was a special onset function, the initial half of a cosine-squared window function, for smoothing the onset of the vowel. This was followed by an 80 msec identity function. Next, the second half of the cosine-squared window was used to smooth the offset of the vowel. Finally, the remaining portion was deleted, multiplying the signal by zero.

Apparatus

A Hewlett-Packard 204D oscillator was used to generate a sine wave of 500 Hz. The stimuli were prepared using the Alligator system and the PDP-12 minicomputer. An audio-frequency filter (Rockland model 1524-01) was used to bandpass-filter the signal between 68 and 6800 Hz. The stimuli were presented on-line to listeners over headphones (Telephonics TDH-49 (frequency response: 30 to 6000 Hz ± 3 dB). A power amplifier (Braun AG Type CSV 250) and an intensity meter (Hewlett-Packard model 3469b) were used to adjust the amplitude of the playback signal.

Procedure

The stimuli were presented with the aid of an Alligator program to listeners over headphones, in a relatively quiet phonetics laboratory. This program carried out the following

steps: each digitized vowel was transfered from disc storage to the Alligator work area, in accordance with a previously generated randomization list. Two such lists were prepared, as in Experiment 4: in the blocked condition vowels were randomized for each speaker individually, while in the mixed condition speakers were randomized as well. The vowel randomizations were the same for both sets; only the identity of the speaker changed. Each vowel appeared once in every set of ten. No two vowels were repeated on adjacent trials. Each vowel was modified by the previously generated window function described above. Each vowel segment was presented following a 5 second inter-stimulus interval. An additional 2 second pause was included after every five vowels.

Each listener completed four sessions in total, including both blocked and mixed conditions in both orders. They were instructed that they would hear ten vowels from a group of speakers. The identity of the vowels was indicated on a handout by means of phonetic transcription symbols. They were asked not to omit any responses or alter their responses to previous tokens. None of the listeners had any difficulty with the task.

Results and Discussion

The overall error rate in the mixed speaker condition was 13.75 percent; in the blocked speaker condition, 9.5 percent. Table 16 presents the mean error rates in each

condition for each of the vowels. To determine the effects of order and practice, each listeners' responses were recorded in each of the four conditions (blocked and mixed speaker conditions, in both orders). Order (mixed condition followed by blocked, or vice versa) was treated as a separate factor. A 2x2 fully repeated measures ANOVA was conducted on the factors order and context (mixed versus blocked), collapsing across vowels and speakers for each of the 8 listeners. Table 17 presents the outcome of this analysis. A significant main effect was obtained for speaker context ($F=10.92, d.f.=1,28; p<.01$). No significant differences were observed for the factor order or for the order-by-speaker context interaction. These findings appear to confirm the hypothesis that dynamic characteristics convey perceptually relevant information for vowel identification. When vowels are artificially shortened, the error rate increases, with more than twice as many errors in each condition. A significant improvement is observed in the blocked condition, indicating that when dynamic information is not present listeners appear to rely on speaker context to disambiguate confusing vowel pairs.

Examining the confusion matrix in Table 18, it can be seen that nearly all of the errors involve confusions between vowels which are adjacent in the vowel diagram. Prominent among such errors are / e - I /, / e - ɛ /, / æ - ɛ /, and / ɒ - ʌ / confusions. All of these substitution errors are markedly reduced in the blocked

Table 16. Experiment 5: vowel identification errors.

	MIXED		BLOCKED	
Listener	M - B	B - M	M - B	B - M
1	14	18	14	11
2	16	12	9	5
3	13	10	8	10
4	19	16	7	7
5	14	9	14	16
6	18	18	10	11
7	17	10	9	5
8	8	8	12	4

Legend:

M: Mixed
B: Blocked

Table 17. Experiment 5: ANOVA

SOURCE	df	Sum of Squares	Mean Square	F
O	1	32.00	32.00	2.42
C	1	144.50	144.50	10.92**
OC	1	0.50	0.50	0.04
L(OC)	28	370.50	13.23	
Total	31	4872.00		

** p .01

Legend:

O: order (Mixed - Blocked vs. Blocked - Mixed)
C: speaker context (vowels presented in Blocked or Mixed condition)
L: listeners

Table 18. Experiment 5: vowel confusion matrix

Vowel Response

Vowel Presented		i	ɪ	e	ɛ	æ	ʌ	ɒ	o	ʊ	u	ø
i	M	61	7	6	2		1			1	2	
	B	67	8	1							4	
ɪ	M	8	64	4	3					1		
	B	1	68	7	3		1					
e	M	6	41	13	19					1		
	B		32	35	12							1
ɛ	M			2	74	4						
	B			1	74	5						
æ	M				22	51	6	1				
	B				14	58	7					1
ʌ	M					6	60	11		3		
	B					9	61	9		1		
ɒ	M						28	52				
	B						10	70				
o	M							2	69	9		
	B						1	1	70	8		
ʊ	M		1	1	1		6			64	7	
	B						3	2	1	70	4	
u	M									8	72	
	B									5	75	

Legend:

M: Mixed speaker condition
B: Blocked speaker condition
ø: no response

condition. A large proportion of the improvement in the blocked condition is due to the vowels / e / and / o /. Small improvements are found for / ʌ, ɪ, æ, ʊ / as well.

Even though gated vowels are identified with higher error rates than full-length isolated vowels, the error rate in the mixed condition is still remarkably low, considering that many of the sources of variation in connected speech have been removed.

This result suggests that vowels carry a great deal of redundant information; several acoustic variables may be serving simultaneously as cues to vowel identity. For example, vowel pairs such as / e - ɛ / are characterized by differences in formant frequencies and in duration, as well as in the presence and direction of diphthongal offglides. It is not known how these features interact in connected speech, but there are indications of context-conditioned differences in the prominence of particular cues. Koopmans-van Beinum (1973) measured the formant frequencies and durations of Dutch vowels under different conditions of speaking rate. She found that duration differences are neutralized in rapid speech. While a comparable study has not been conducted for the English vowels, it seems likely that similar results would be obtained. Peterson and Coxe (1953) also found that diphthongization of vowels varies with prosodic and consonant context.

Vowels in isolation or CVC context are redundantly specified by a number of acoustic parameters, each of which

may contribute to the recognition of the vowels. Under special conditions, such as rapid speech, some of these acoustic properties may be neutralized. Under these circumstances one might expect factors like speaker information to play a more significant role. Careful measurement studies, combined with perceptual experiments are needed to determine whether listeners use this information.

It has been assumed throughout that the data utilized in this study show considerable overlap in their formant frequencies, and that perceptually this overlap is responsible for vowel confusions.

A series of measurements were taken on the isolated vowel segments used in Experiment 5 to determine if confusion errors are related in any simple way to formant overlap.

B. Acoustic analyses of gated vowels

A series of quantitative analyses were conducted on the data used in the gated vowel study. Under the assumption that information present in the signal is a primary determinant of the identification responses of listeners, it was anticipated that acoustic analyses would help to account for the results obtained in the perceptual experiment.

The vowel segments analyzed were 80 msecs in duration. It is assumed that a single time-slice of the vowel contains essential information for its identification; and that dynamic characteristics such as duration and diphthongization are suppressed or eliminated by the gating procedure.

Earlier studies have indicated that the essential physical specification of a vowel is provided by information given by the frequencies of f0, F1, F2, and F3. While higher formants may contribute to greater naturalness, there is little evidence that thay contribute to vowel quality.

According to a relative formant normalization hypothesis, overlap in the formant space should lead to an increase in identification errors, for context-free vowels; speaker context should help to reduce the ambiguity of acoustically similar tokens. Moreover, vowels which are "representative" in terms of their formant frequencies should receive high recognition scores.

This leads to two interesting predictions:

1. formant overlap should be reflected in an increase in errors, at least for context-free vowels.
2. non-overlapping tokens should be well-recognized under all conditions.

The use of phonetic judgements offers several additional hypotheses. Phoneticians' judgments of height and advancement are related to F1 and F2 values (Joos, 1948). The perceptual basis for this correlation has not been

investigated in detail. In addition, no previous study has examined the effects of speaker-dependent variations in formant frequencies on phonetic judgements. If phoneticians rely on information provided by the formant frequencies of vowels, how do they compensate for variations due to context? It is not known whether phoneticians' judgments are related to identification responses by listeners.

The statistical methods of linear discriminant function analysis are adopted to determine the distribution and distinctness of the vowels within the acoustic and auditory space. Comparisons are made between the two domains and the perceptual data of Experiment 5.

Stimulus materials

Each of the 100 isolated vowels was modified by a window function similar to the one described in Experiment 5.

Apparatus

Acoustic analysis was conducted using the PDP-12. Vowels were accessed from Alligator disc storage by means of an OS/8-Alligator interfacing program.

Procedure

Acoustic analysis was conducted by means of programs described in Nearey, Hogan and Rozsypal (forthcoming). The spectral analysis proceeded as follows: Each 80 millisecond

segment was multiplied by a Hamming window function. The windowed signal was then subjected to an autocorrelation LPC analysis using Markel and Gray's subroutine AUTO (1976:219) with 20 predictor coefficients. A smoothed estimated spectrum of the signal was derived from the predictor coefficients in the manner described by Markel and Gray (p.160f). Formant candidates were selected by a technique based on that described by Christensen, Strong and Palmer (1976). The six candidates with the highest amplitudes were selected. The four lowest frequency candidates thus selected were assigned to formants one to four in order of increasing frequency. f0 was estimated by a cepstral analysis (Noll, 1963). Inspection of the resulting measurements indicated a small number of gross errors (missed and/or spurious formants). These values were corrected manually. The data are presented in Table 19.

Statistical analysis

Raw formant measurements were converted to (natural) log values, for reasons discussed in Nearey (1977: 221). Formant frequencies have been shown to exhibit a linear correlation between means and standard deviations; log-transformed values do not show this correlation.

The following notation will be used:

G0 = log f0;

G1 = log F1;

G2 = log F2;

Table 19. Formant frequencies (hertz)

		T1	T2	T3	T4	T5	T6	T7	T8	T9	T10
i	F0	113	91	121	89	89	192	180	222	191	235
	F1	301	285	254	332	231	371	356	263	379	465
	F2	2426	2106	2371	1941	2145	2504	2691	2418	2660	2446
	F3	3262	2637	3006	2934	2895	3363	3434	2926	3410	3066
	F4	3663	3363	3941	3473	3441	3832	4020	3527	4191	3512
ɪ	F0	125	91	122	84	91	188	205	255	189	229
	F1	441	410	387	481	426	457	449	566	449	512
	F2	2160	1801	2223	1723	1926	2316	2340	2410	2254	2269
	F3	2637	2410	2621	2668	2535	2938	2934	2629	3035	2926
	F4	3715	3230	3824	3473	3457	3965	3996	3316	4137	3691
e	F0	114	44	114	86	81	182	176	100	182	112
	F1	566	473	457	473	402	535	504	559	480	483
	F2	2207	1832	2191	1777	2121	2340	2527	2566	2355	2223
	F3	2801	2418	2629	2613	2645	2996	3379	3483	2996	2871
	F4	3746	3238	3879	3465	3512	3930	4020	4270	4184	3348
ɛ	F0	110	87	114	83	80	184	182	213	184	222
	F1	645	652	590	559	574	660	582	684	613	676
	F2	1941	1621	2027	1676	1793	2160	2340	2340	2191	2137
	F3	2598	2285	2543	2520	2496	2957	2957	3262	2949	2809
	F4	3707	3301	3918	3457	3613	3941	3988	3691	4199	3506
æ	F0	115	86	110	79	78	191	176	191	183	220
	F1	770	801	793	645	730	941	762	957	801	348
	F2	1715	1457	1691	1605	1660	1957	2191	1941	1777	1699
	F3	2480	2277	2316	2520	2332	2660	2762	2824	2316	2668
	F4	3676	3660	3879	3598	3504	3715	3770	3996	3395	3605
ʌ	F0	118	105	115	81	83	182	184	203	177	216
	F1	809	770	723	746	746	879	840	980	824	824
	F2	1324	1285	1270	1285	1191	1254	1324	1613	1418	1293
	F3	2496	2152	2043	2309	2160	2793	2994	2871	2895	2590
	F4	3512	3293	3582	3504	3434	4035	4152	3996	3957	3629
ɒ	F0	116	89	110	80	83	180	185	201	181	221
	F1	738	723	770	660	691	809	652	809	738	777
	F2	1043	996	1090	1051	1012	1184	1098	1184	1199	1082
	F3	2402	2449	2020	2223	2402	2746	2746	3051	2801	2629
	F4	3387	3223	3441	3184	3387	3301	3855	4160	3809	3016
o	F0	122	91	115	84	84	177	177	210	180	223
	F1	551	543	481	449	395	551	504	566	480	660
	F2	980	895	816	879	746	988	996	1043	934	1059
	F3	2566	2277	1926	2215	2254	2668	2598	2926	2676	2793
	F4	3637	2777	3629	2941	3051	3574	3481	3973	3840	3777
ʊ	F0	115	93	112	89	81	178	194	221	178	222
	F1	449	457	457	457	465	543	590	590	473	598
	F2	1066	1176	1105	1301	1426	1285	1293	1738	1527	1270
	F3	2520	2246	2144	2301	2238	2660	2519	2863	2762	2676
	F4	3621	3098	3628	3379	3480	3621	3488	4309	3824	2777
u	F0	121	92	124	86	96	185	186	245	183	228
	F1	355	356	277	348	270	363	363	355	363	449
	F2	1035	910	981	1207	941	1093	1145	1926	1093	1056
	F3	2340	2333	2160	2035	2207	2559	2556	2668	2566	2504
	F4	3652	3027	3879	3059	2973	3707	3715	3793	3981	3613

G3 = log F3.

Formant overlap between vowels across speakers is illustrated graphically in Figure 2 (G1 by G2 for all 10 vowels for all 10 speakers) and Figure 3 (G2 by G3 for all vowels). Overlap is particularly severe between the vowels / e - I / and / ʌ - ʊ /. However, it is difficult to demonstrate graphically what independent contribution each of G0, G1, G2, G3 and G4 make to keeping the vowels distinct. A statistical method, linear discriminant function analysis, was adopted to this end. This procedure assigns each token to the group (vowel category) for which its calculated probability of membership is highest relative to other groups, based on a set of measured variables (formant frequencies and fundamental frequency). The decision procedure will assign a token to the "correct" category (ie. the group to which it actually belongs) if the calculated probability of "correct" group membership (henceforth P(G|x)) is higher than the probability of membership in any other group. P(G|x) provides a statistical estimate of the probability that a token with a given set of formant frequencies is a member of the "correct" vowel category (or the group from which it is drawn). Formally, this measure can be defined as follows:

$$P(G_j|X) = Gau(X,M_j,C)\cdot B_j / \sum_{i=1}^{n} Gau(X,M_i,C)\cdot B$$

where:

B_j = a priori probability of membership in group j;

X = vector of measured values

M_j = mean vector for jth vowel;

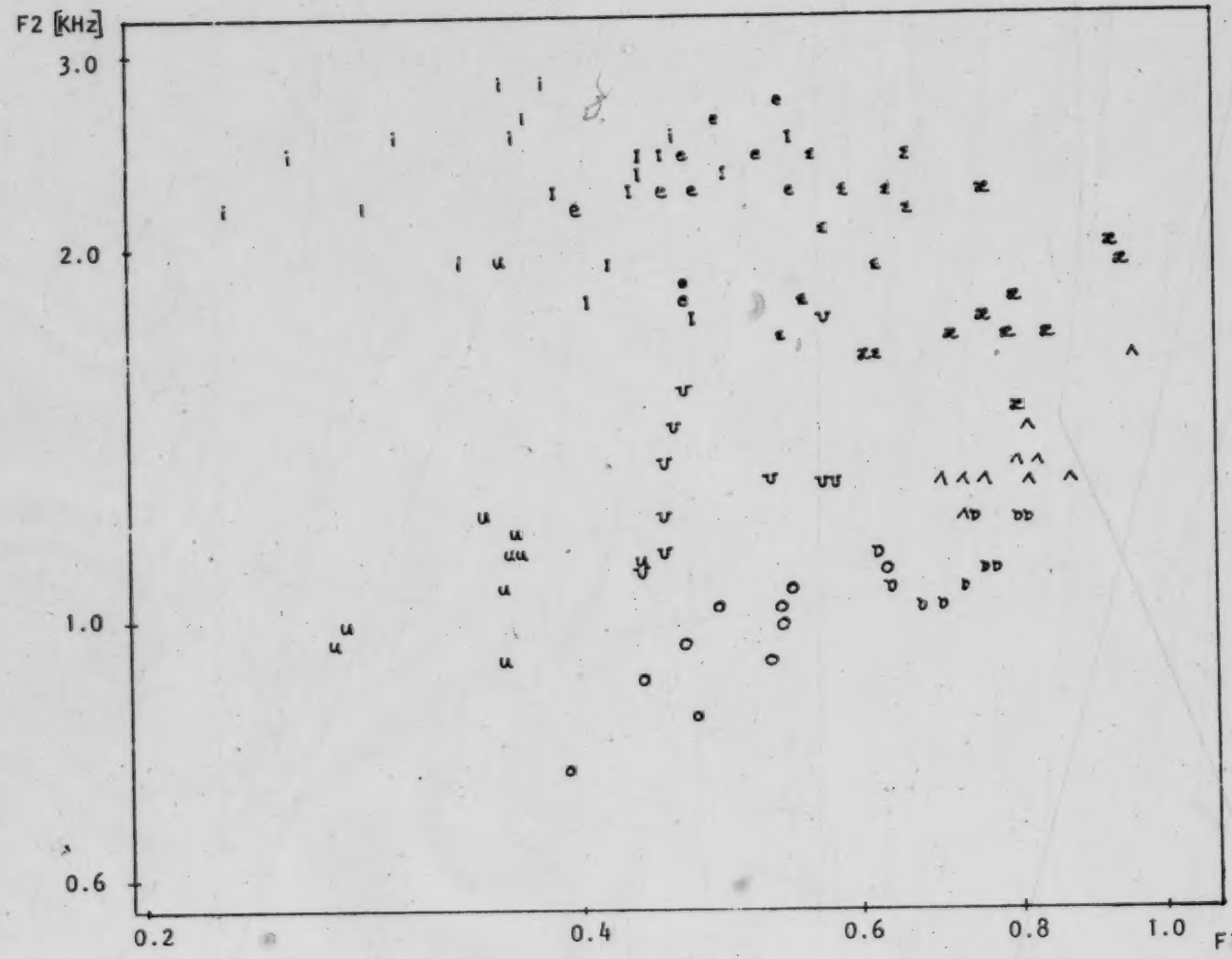


Figure 2. Frequencies of F1 and F2 [KHz] in log scale

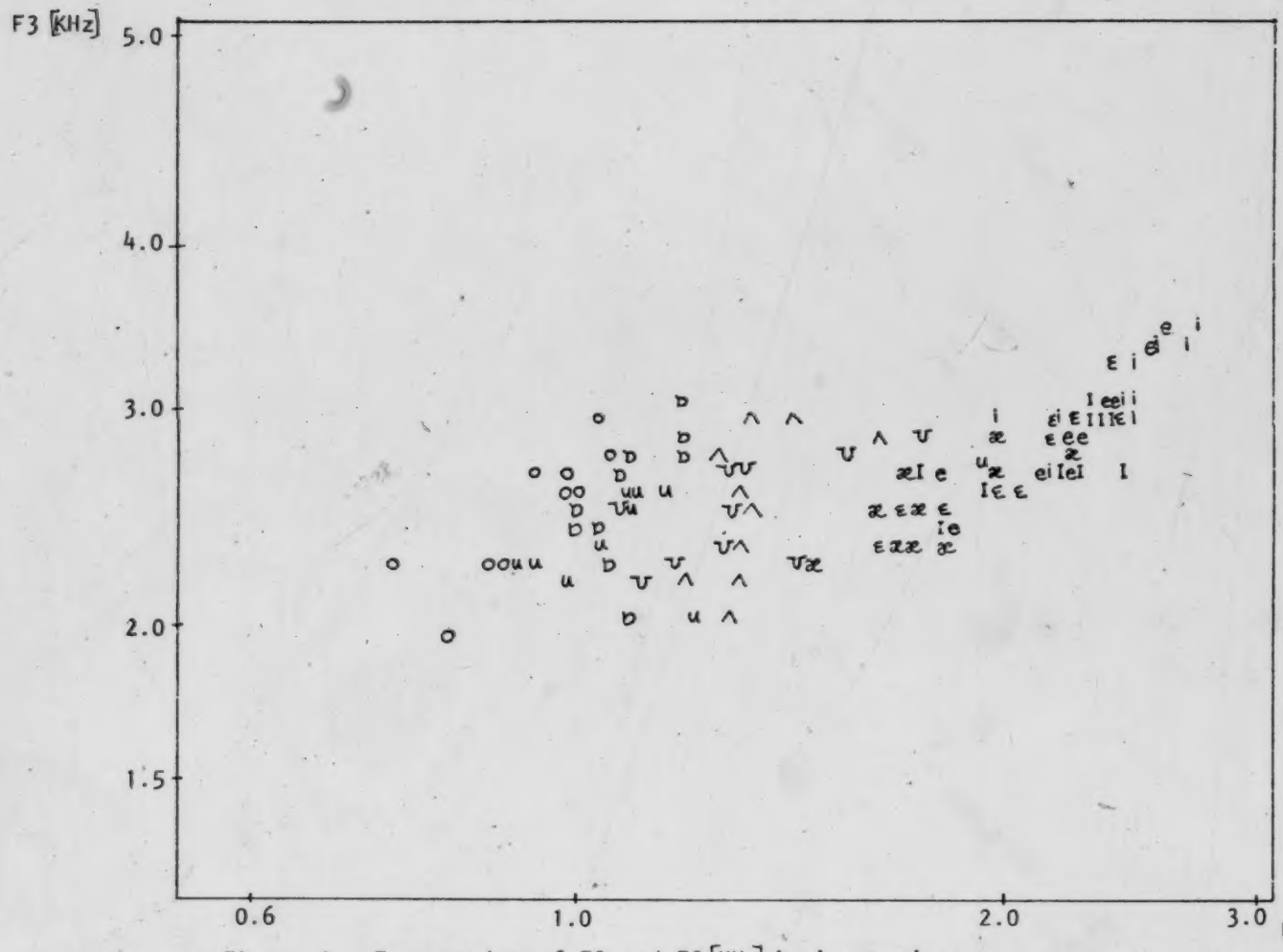


Figure 3. Frequencies of F2 and F3 [KHz] in log scale

C = pooled-within-groups covariance matrix

Gau = multivariate Gaussian probability density function.

When the subscript j is not specified, it will be assumed that G refers to the "correct" or "intended" vowel category.

Equal variance-covariance matrices are used and a multivariate normal distribution is assumed. It is assumed that variance-covariance matrices are equal in the population from which the sample is drawn, and that groups differ only in their means. In all of the analyses described below, classification was based on two discriminant functions. Additional functions contributed only marginally to the overall solution. The direct method was used with no rotations. A priori probabilities were assumed to be equal in preliminary analyses. Further information is given in Tatsuoka (1970) and Nie, Hull, Jenkins, Steinbronner and Bent (1975).

Discriminant analysis was used to obtain a statistical indication of the "relative discriminability" of the vowels in terms of their formant frequencies and fundamental frequency. Two measures of "relative discriminability" were used. The first measure is the overall percent correct classification, or the percentage of tokens assigned to the groups to which they actually belong. The second measure is the average P(G|x) (henceforth P(G|x)). P(G|x) is the calculated probability of correct group membership, as described above. P(G|x) therefore provides an index of how

representative the tokens are of the vowel classes from which they are drawn. P(G|x) and percent correct classification values were obtained from the SPSS package program for statistical analysis (Nie et. al.,1975).

Both of these measures can be used to reflect the degree of overlap between vowel categories in the formant space. Large values (to a maximum of 1.0) will indicate good separation of the tokens with respect to the vowel categories. Smaller values indicate extensive formant overlap among the vowels, or poor separation of tokens with respect to vowel categories.

For example, if a token whose formant frequencies lie near the mean of the distribution for the intended vowel class, it will be correctly assigned to that category; and it will be assigned a P(G|x) value approaching 1.0. As the formant frequencies deviate from the mean of the intended vowel and approach the mean for another vowel, the P(G|x) value will decrease. In some cases formant frequencies of individual tokens will deviate substantially from the mean values of the intended vowel class. Still, these values may be even more remote from the means for other vowels. The P(G|x) score will reflect this, and the token will be assigned to the correct category. Classification errors will result when the probability of group membership is higher for a vowel class other than the intended category.

What is the degree of formant overlap in the vowels used in the gated vowel study? Table 20 summarizes the

Table 20: statistical discrimination of vowels

Parameters	Discriminant analysis: % correctly classified	Average P(G/x)
G1, G2	83	.7209
G0, G1, G2	89	.8100
G1, G2, G3	81	.7657
G0, G1, G2, G3	89	.8084
CLIH: NG1, NG2	87	.8179
CLIH2: NG1, NG2	93	.8566
CLIH3: NG1, NG2, NG3	96	.8635
CLIH2 with G0: G0, NG1, NG2	94	.8574
Phonetic judgements: height & advancement	95	.8957

Table 21: correlations (Pearson r) between P(G/x) scores and identification rates for individual tokens

Parameters	% correctly identified: Blocked condition	% correctly identified: Mixed condition
G1, G2	.428	.252
G0, G1, G2	.429	.176
G1, G2, G3	.3516	.259
G0, G1, G2, G3	.424	.180
CLIH: NG1, NG2	.339	.226
CLIH2: NG1, NG2	.321	.276
CLIH3: NG1, NG2, NG3	.337	.300
CLIH2 with G0: G0, NG1, NG2	.316	.268
Phonetic judgements: height & advancement	.290	.107

output from a series of linear discriminant analyses conducted on the formant data. On the basis of information in G1 and G2 alone, 83 percent of the tokens are correctly identified by the discriminant analysis procedure. When G0 is added, identification increases to 89 percent. When G3 is included with G0, G1, and G2 there is apparently no improvement. When G3 is added to G1 and G2 the classification rate actually goes down (81 percent). These results suggest that G1 and G2 can discriminate reliably among the vowels; and that G0 makes a significant contribution. However, these results suggest that G3 does not play an important role.

The second measure of statistical discriminability, the average $P(G|x)$, indicates that: a) G1 and G2 alone discriminate successfully among the vowels ($P(G|x)=.7209$); and b) G3 may be somewhat more important if G0 is unavailable. An increase (from .7209 to .7657) is found when G3 is included with G1 and G2 alone. However, a slight decrease is observed in the average $P(G|x)$ value if G3 is included with G0, G1, and G2 (from .8100 to .8084). Once again, statistical evidence indicates a relatively minor role for G3 in discriminating between the vowels.

These results are generally in agreement with the overall identification rates obtained in the perceptual studies. The overall error rate of 13.75 percent obtained for context-free vowel segments (gated vowels in the mixed condition) is only slightly higher than the

misclassification rate of the statistical procedure.

The second half of Table 20 presents the results of a second series of analyses, using speaker-normalized formant measures. The normalization procedures were adopted from Nearey (1977). The first procedure is based on CLIH (constant log interval hypothesis). According to this hypothesis, the formant frequencies of different speakers are related by a single speaker-dependent parameter.

The normalization procedure thus involves the subtraction of a single parameter from the raw formant values (in the log scale) of each vowel token. This parameter is the average value of G1 and G2 combined, for all of a speaker's vowels.

NG1, NG2, and NG3 will be used to represent normalized values of G1, G2 and G3, respectively, under a given normalization procedure. G1AV, G2AV, and G3AV will be used to denote the average values of G1, G2, and G3, respectively, for all of a speaker's vowels. Thus, for example:

$$G1AV = 1/n \sum_{i=1}^{n} G1$$

where G1 is a given speaker's G1 for vowel i; and n is the number of vowels under consideration.

CLIH:

NG1 = G1 - ((G1AV + G2AV)/2)

NG2 = G2 - ((G1AV + G2AV)/2)

CLIH2 involves the subtraction of two speaker-dependent parameters. Separate averages for each formant are

calculated for a given speaker's vowels. These values are subtracted from the raw formant values (in log scale) for each token.

CLIH2:

NG1 = G1 - G1AV

NG2 = G2 - G2AV

CLIH3 is identical to CLIH2, but involves an additional third parameter for G3.

CLIH3:

NG1 = G1 - G1AV

NG2 = G2 - G2AV

NG3 = G3 - G3AV

Discriminant analyses were conducted on the formant data as modifed by each of the three normalization schemes. The inclusion of speaker-dependent information improves the overall classification rate. CLIH results in a 4 percent improvement over raw G1 and G2 in terms of the number of correctly classified cases (87 percent). However, the average P(G|x) increases by almost 10 percent (.8179). CLIH2 raises the classification rate to 93 percent, and the average P(G|x) value to .8566. CLIH3 provides the "best" model, in terms of the overall separability of the vowels; classification rate is 96 percent, and average P(G|x) is .8635. However, it is interesting to note that discriminant analysis with the parameters of CLIH2 plus G0 included as an additional parameter obtains comparable results.

If the improvement in the blocked condition is due to

perceptual normalization of the sort described by relative formant normalization procedures, it should be possible to test such schemes in terms of their efficacy in accounting for the perceptual data. If speaker information (as implemented by a given normalization procedure) helps to render ambiguous vowels more distinctive, the presence of speaker information in a perceptual test should aid the listener to identify the vowel token.

Measures of overall average discriminability of the vowels may be used as a crude measure of the information potentially available; yet it would be more useful to compare each individual vowel in terms of a) its position in the formant space; and b) identification responses by listeners. On the assumptions of a relative formant normalization model, vowels which show large deviations in their formant frequencies (resulting in overlap across speakers) should be frequently misidentified by listeners. Similarly, tokens which are highly representative in terms of their formant frequencies should be well-identified by listeners.

The measure $P(G|x)$ for <u>individual tokens</u> offers a method for making comparisons of this sort. This measure may be used as an index of "strength of group membership" of a given token with respect to the formant frequencies of the vowel which it represents, relative to other groups. If identification responses made by listeners depend on the "representativeness" of a token in terms of its formant

frequencies, a significant linear correlation might be predicted between the P(G|x) value for a given token and the proportion of correct identifications by listeners. High correlations are to be expected if the parameters selected contain a complete (or nearly complete) representation of the information upon which listeners base their identifications. Low correlations will result if the listener makes use of additional information provided in the signal or its context, or if he is unable to extract the information utilized by the statistical procedure.

The improvement in the blocked condition over the mixed condition may be due to the presence of speaker information. If so, the increase in the identification rate should parallel the improvement when speaker information is added in the statistical procedure. If listeners have access to this information, the P(G|x) scores for individual tokens should be correlated with the identification scores.

In the mixed condition speaker context is unavailable. Error rates should therefore correlate more strongly with raw formant measures.

Table 21 presents correlation coefficients (Pearson r) for the P(G|x) score of each token with its percent correct identification rate (by listeners). The correlations are all in the predicted direction. Nearly all are statistically significant. However, the differential predictions concerning the role of speaker information are not borne out. The highest correlations occur between raw formant

measures and identification scores in the blocked condition. Normalization appears to decrease the correlation with the blocked condition, while in some cases improving the correlation with identification scores in the mixed condition. For example, P(G|x) scores for the CLIH-normalized data show correlations with identification rates in the blocked condition at r=.339. Yet P(G|x) scores for raw G1 and G2 measures correlate with the blocked data at r=.428, a substantial increase.

Since the stimuli in the perceptual study were gated vowels, it is possible that there were response biases present which decrease the correlation between the two variables. Listeners may tend to respond more frequently with short vowel alternatives. In the present study this may be illustrated by the large number of / e - I / confusions. / I - e / confusions are rare in both mixed and blocked conditions. A second set of discriminant analyses was conducted using adjusted prior probabilities for each of the groups (vowel categories). The prior probabilities were adjusted in accordance with the frequency with which each vowel occurred as a response, averaged across both mixed and blocked conditions.

Table 22 presents a new set of analyses. P(G|x) scores in this set are obtained from discriminant analyses with prior probabilities adjusted for the relative frequency of occurrence of each vowel as a response in the perceptual task; values were averaged across both mixed and blocked

Table 22: correlations (Pearson r) between P(G/x) scores (based on discriminant analysis with prior probabilities) and identification rates for individual tokens

Parameters	% correctly identified: Blocked condition	% correctly identified: Mixed condition
G1, G2	.502	.315
G0, G1, G2	.490	.230
G1, G2, G3	.425	.327
G0, G1, G2, G3	.488	.237
CLIH: NG1, NG2	.416	.296
CLIH2: NG1, NG2	.405	.351
CLIH3: NG1, NG2, NG3	.421	.377
CLIH2 with G0: G0, NG1, NG2	.400	.347
Phonetic judgements: height & advancement	.333	.103

conditions. All of the correlation coefficients increase, indicating that biases may be present. However, the relative differences between the conditions are unchanged. P(G|x) scores for raw G1 and G2 still give the highest correlation with the blocked condition (r=.502). Examination of the data indicates that a large proportion of the tokens which receive high P(G|x) scores are also well recognized by listeners. A number of tokens are well recognized but receive low P(G|x) values. A comparison was made of G1 and G2, and CLIH-normalized G1 and G2 with identification rates in the mixed and blocked conditions. One possible explanation for smaller correlations in the mixed condition is that listeners may be influenced by tokens occurring earlier in the sequence; since the speaker is constantly changing, contextual influences may be negative, resulting in more frequent identification errors. In order to eliminate the possibility of token-to-token contrast effects it would be necessary to conduct the same experiment under different randomizations, and in doing so counterbalance for serial context effects between adjacent stimuli. Effects of this sort are known to occur with synthetic vowels (Fry, Abramson, Eimas and Liberman, 1960). To obtain a reliable estimate of the relative intelligibility of a vowel it may be necessary to obtain a larger data sample, with increased numbers of speakers and listeners.

Several additional factors may be involved. One possibility is measurement error, which would tend to

decrease the correlation between acoustic and perceptual data. Alternatively, listeners may rely on other parameters than those used in the analyses. It is also possible that the assumption of equal variance-covariance matrices is not met in these data. Quadratic discriminant function analysis offers an alternative model which does not depend on this assumption. Additional research is necessary to determine whether these factors are important ones..

However, P(G|x) scores show substantial correlations with the perceptual data, which indicates that formant overlap may be an important factor in determining identification errors.

C. Phoneticians' transcriptions of gated vowels

The perceptual basis for the correlation of F1 and F2 with height and advancement has not been investigated systematically. An auditory analysis of the vowels, based on phoneticians' judgements of vowel features such as height and advancement, may be regarded as a measure of the information present (Nearey, 1977). It has been shown that detailed phonetic analyses can yield insights into the nature of vowel quality. The judgements on which such analyses are based show a high degree of reliability across listeners trained in the same phonetic tradition. (Laver,

1965; Ladefoged, 1967). If height and advancement judgements are determined by variations in F1 and F2, as Joos' findings suggest, they may also reflect the degree of overlap between vowels across speakers. This speculation generates a number of questions:

1. Do judgements of height and advancement show overlap in cases where the vowels are confused by listeners?
2. Do these judgements correlate best with F1 and F2 or do the additional parameters of f0 and F3 also play a role?
3. Do phoneticians use speaker information in their judgements of vowel quality?

It was anticipated that answers to some of these questions might be useful in explaining the findings obtained in the perceptual experiment.

The method of discriminant function analysis was again used to determine the distribution of the vowels within the height-advancement plane.

Phoneticians

Transcription of the vowels was conducted by two trained phoneticians. One of the phoneticians was a Canadian from the Maritimes. The second was American, a native of New Jersey. Both were highly familiar with the dialect of the speakers.

Stimulus materials

The 100 gated vowels were identical to those used in

the perceptual experiment. The vowels were not randomized but presented in a fixed order:

/ i , ɪ , e , ɛ , æ , ʌ , ɒ , o , ʊ , u /

Speakers were presented in the blocked condition.

Apparatus

The Alligator system was used in conjunction with the PDP-12 to generate the tape used in the transcription task. Presentation was over a HECO loudspeaker using a TEAC tape recorder and a SONY amplifier, as in Experiment 2.

Procedure

The vowels were presented at a comfortable listening level. The phoneticians had full control over the tape recorder, analyzing each token in detail after listening to it several times. Each judgement was based on a consensus between the two listeners.

For each vowel, one of ten phonetic symbols was assigned, with optional diacritics (fronted;'>', retracted;'<', raised;'∧', lowered'∨'). In addition, each vowel was assigned to 2 coordinates: one for height and one for advancement. The coordinates were given by a vowel diagram, partitioned into 80 individual cells (see Appendix 2). Key points were assigned for each of the 10 vowels. Three additional points were specifed for the vowels / ɣ /, / ɯ / and / ɔ /. Table 24 presents the results of the transcription task in terms of the (numerical) position

assigned to each token. It is apparent that while the indicated values for the key symbols may have influenced phoneticians' judgements, they did not serve as an anchor or standard, since in some cases the assigned means are far from the key positions.

Statistical analysis

A discriminant function analysis was conducted on the height/advancement judgements. From Table 20 it can be seen that these features discriminate between the vowels with high accuracy (95 percent correct). The average P(G|x) score is likewise very high (.8957).

Comparisons of P(G|x) scores for individual tokens showed a correlation of r=.290 with identification responses in the blocked condition, and r=.107 in the mixed condition. An additional test was conducted using prior probabilities based on the average frequency of occurrence of each vowel, as in the acoustic data. This analysis yielded slightly higher correlations in the blocked condition (r=.333) but little change in the mixed condition (r=.103). For both analyses only the correlation in the blocked condition was statistically significant ($p<.01$). Since the vowels were transcribed under a blocked speaker condition, it is possible that speaker information may account for this difference.

Table 23. Multiple Regression: phonetic judgements with formant measures

Variable	Controlling for:	R^2 change	df	F	R^2
		I. Height			
G1	----	.7735	1,98	334.721**	.7735
G0,G2,G3	G1	.0660	3,95	13.0218**	.8395
G1AV	G1,G0,G2,G3	.0093	1,94	5.782*	.8488
G1AV	G1	.0630	1,97	37.3761**	.8365
G0,G2,G3	G1,G1AV	.0123	3,94	2.5489	.8488
		II. Advancement			
G2	----	.9032	1,98	914.611**	.9032
G0,G1,G3	G2	.0365	3,95	19.168**	.9397
G2AV	G2,G0,G1,G3	.0042	1,94	7.0374**	.9439
G2AV	G2	.0334	1,97	51.1009**	.9366
G0,G1,G3	G2,G2AV	.0073	3,94	4.0772*	.9439

* $p<.05$
** $p<.01$

Correlations with acoustic data

Statistically reliable information is provided by the inclusion of one or more speaker dependent parameters along with raw formant measures. It was hypothesized that (P(G|x) scores based on a discriminant analysis of normalized formant values would correlate best with identification scores in the blocked condition. This hypothesis was not supported.

Phoneticians' judgements offer an additional test of the normalization hypothesis. While the correlation of these judgements with raw formant measures is reported to be strong, the question which arises is whether the inclusion of speaker information can improve this correlation. A series of heirarchical multiple regressions were conducted to answer the following questions:

1. Is height (advancement) significantly correlated with G1 (G2) in the absence of additional parameters?
2. Does the inclusion of a speaker dependent parameter, G1AV (or G2AV), the mean value of G1 (G2) for all of a given speaker's vowels, significantly improve the correlation with height (advancement) when other sources of information in the signal (such as G0, G2 (G1) and G3) are controlled for?
3. When the effects of G1AV (G2AV) are controlled for, do G0, G2 (G1), and G3 still make a contribution to the overall correlation with phoneticians' judgements?

Judged height (advancement) was used as the dependent variable. G1 (G2); G0, G2 (G1), G3; and G1AV (G2AV) were used as predictor variables. Table 23 presents the results of this analysis. Correlations between height and G1 are high (r=.880). When the speaker parameter G1AV is included a significant improvement in is found (F= 37.376; d.f.=1,97; p<.01). Controlling for the addition of the parameters G0, G1 and G3, G1AV still contributes significantly to the correlation between G1 and height (F=5.78; d.f.=1,97; p<.05).

When G0, G2 and G3 are simultaneously entered into the equation and G1AV is controlled for, no significant improvement is found. Separate analyses were conducted for all possible combinations of these three factors. G0 and G2 together make a significant contribution (F=3.766; d.f.=2,95; p<.05), as do G0 and G3 together (F=3.441; d.f.=2,95; p<.05). Taken separately, however, only G0 results in a significant improvement (F=4.959; d.f.=1,96; p<.05).

Similar results are obtained for correlations between G2 and advancement. These factors are highly correlated (r=.95). Addition of the speaker-dependent parameter G2AV results in a significant improvement (F=57.101; d.f.=1,97; p<.01). Controlling for G0, G1 and G3, G2AV still makes a significant contribution (F=7.037; d.f.=1,94; p<.01).

G0, G1 and G3 taken together contribute significantly to the correlation (F= 19.168; d.f.=3,95; p<.01). Controlling

Table 24: phonetic judgements of vowel segments

	i		I		e		ɛ		æ		ʌ		ɒ		o		U		u	
S	H	A	H	A	H	A	H	A	H	A	H	A	H	A	H	A	H	A	H	A
1	.5	1	1	2	3	2	5	3	7	4	6	4.5	6	7	1.5	7	1	6	.5	7.5
2	.5	1	1	2	3	1.5	6	3	8	4	5.5	5	7	6	3	7	1	5.5	.5	8
3	.5	.5	1	2	2.5	1.5	4	2	7.5	4	6	5.5	6	6	2.5	7.5	1	5	1.5	8
4	0	1	1	2	2	1	4	2	5.5	3	5.5	5.5	6	6.5	2	8	2	5	0	7
5	.5	0	1.5	1.5	2.5	2.5	4	2	7	3.5	5.5	5.5	5.5	6.5	2	8	1	5.5	0	7.5
6	1	1	1.5	2.5	1.5	1	4	2	7	4	6	4.5	6	7	3	8	1	5	.5	7
7	0	0	1	2	2	1.5	4.5	2	7.5	3	6	4	5	6.5	2	7	1	4	0	4
8	.5	0	1	2	2	1	4	2	7	4	5	5.5	5	6.5	2	7	1	4.5	0	7.5
9	1	1	2	2	2.5	1.5	4	3	6	3	6	5.5	6	6.5	2.5	7.5	1	6	0	7.5
10	.5	.5	1	2	2	1	4	2	7	3.5	6	5	6	6.5	1.5	8	1	4.5	0	7.5

for G2AV, the correlation is still significant, though reduced ($F=4.077$; $d.f.=3,94$; $p<.05$). Separate analyses were conducted on all possible combinations of these three variables. The only pairwise combination resulting in a significant improvement was G0 and G3 ($F=6.202$; $d.f.=2,95$; $p<.05$). Taken separately, statistical significance was obtained for G0 ($F=6.074$; $d.f.=1,96$; $p<.05$) and G3 ($F=4.007$; $d.f.=1,96$; $p<.05$).

The normalization procedure CLIH2 subtracts from each G1 or G2 score the subject mean for that formant across the entire set of his vowels. In the regression, however, an optimal weighting for the parameter G1AV or G2AV is determined in order to maximize the overall correlation with the dependent variable height (advancement).

When the variables G1 and G1AV are combined optimally to predict height judgments, the B-weights assigned are approximately inversely related to one another, which is the relationship predicted by CLIH2 (B= 6.975 for G1, -6.944 for G1AV). In the case of G2 and G2AV there is a slight deviation from CLIH2 (B=-7.054 for G2, and 5.042 for G2AV).

Speaker information therefore appears to account for a significant proportion of the variations in height/advancement judgements, even when the intrinsic vowel parameters of G0, G3 and G1 (or G2) are taken into account. It is interesting that these additional parameters also make a significant independent contribution.

Since CLIH2 offers a close approximation to the

regression solution, further correlations were computed between CLIH2-normalized G1 and G2 with height and advancement. Highly significant correlations were obtained. Correlations between height and CLIH2-normalized G1 (r=.9146) were slightly lower than the correlations between advancement and CLIH2-normalized G2 (r=.9648).

One objection which might be raised is that this correlation is to some extent predetermined by the nature of the vowel diagram, and the placement of key positions for each of the vowels. One method to correct for this factor involves subtracting the mean for each vowel from the normalized formant values for each token and from the mean value assigned to the corresponding height or advancement judgements. Correlation coefficients were calculated between the normalized formant values, expressed as deviations from the vowel means , and height/advancement judgements, also converted to deviation scores from the vowel means. Even under this condition a significant correlation was obtained for the transformed variables (r=.259 for transformed height and normalized G1; r=.535 for transformed advancement and G2).

In summary, the findings obtained suggest a role for speaker parameters not radically different from those provided by CLIH2. In addition, they suggest that fundamental frequency and remaining formants are correlated with phonetic judgements as well. Further studies, perhaps involving synthetic speech, are necessary to determine

whether these parameters constitute the perceptual basis for judgments of height and advancement.

V. SUMMARY AND CONCLUSIONS

In Chapter 1 the theoretical motivation for the experimental investigation of context are discussed. Central to this discussion is the finding of formant frequency variations in vowels. Two factors, speaker differences and consonant context variations, have received special attention in the literature on vowel perception.

A review of the literature indicates a large discrepancy in the experimental results. In some experiments consonant environment appears to have had a large effect on vowel identification; in others, context-free vowels are reported to be well identified.

Some possible reasons for these differences in results are proposed. Labeling difficulties and spelling confusions may affect the responses of listeners if the task is a written one. An experiment is described in which listeners were asked to give both written and spoken (repetition) responses. Spoken responses were later transcribed by phonetically trained listeners. Vowels are well identified in the spoken condition, but frequently labeled incorrectly in the written condition.

A second experiment compares the effects of different response frames on error rates, for the identification of isolated vowels and / pVp / syllables. This experiment

indicates that some of the reported improvement due to consonant environment may be artifactual. When spelling responses are required, listeners obtain very high identification scores. Keyword responses (written responses which are different from the syllable presented) appear to be more difficult, and generate more errors.

This problem may be avoided by using one of three types of responses in vowel identification experiments:

1. oral repetition responses, which are later transcribed by phonetically trained listeners;
2. keyword responses with a moderate amount of training;
3. phonetic transcription by trained listeners.

The latter two methods are used in Chapter 3 to evaluate the effects of consonant and speaker context. These experiments suggest that both types of context make a significant contribution, but may be less important in vowel perception than previously indicated. It is suggested that dynamic characteristics such as duration and diphthongization differences among the vowels may provide additional cues for vowel identification. Evidence for this view is presented in a study using gated vowels. Errors are more frequent, and listeners appear to benefit from the presence of speaker context.

Acoustic analyses are conducted on the vowel segments used in the gated vowel study. Discriminant analysis of this data indicates that there is sufficient information contained in the steady state segments to discriminate

between the majority of the vowels. When one or more speaker dependent parameters are added, the classification score is even higher (up to 96 percent). Correlations with perceptual data indicate that while identification responses are closely related to formant frequency variations, improvements due to context in listeners' identifications of vowels apparently do not depend on the speaker parameters used in the statistical analyses.

Detailed phonetic transcription of the gated vowels were also obtained. Very strong correlations exist between formant frequency values and phonetic judgements of height and advancement. Statistical analysis of height and advancement judgements indicates that speaker information strengthens the correlations with raw formant measures.

The results reported in this study indicate that vowel perception is generally very robust. Context may play a role under certain conditions. However, it appears that the magnitude of certain kinds of context effects may have been overestimated in previous research.

REFERENCES

Ainsworth, W.A. 1972a. Duration as a cue in the recognition of synthetic vowels. Jour. Acoust. Soc. Am. 51: 648-651.

Avis, W.S. 1975. The phonemic segments of an Edmonton dialect. In Chambers, J.K. (ed.) Canadian English: origins and structures Toronto: Methuen.

Bennet, D.C. 1968. Spectral form and duration cues in the recognition of English and German vowels. Language and Speech 11: 65-85.

Carlson, R., Granstrom, B., and Fant, G. 1970. Some studies concerning the perception of isolated vowels. STL-QPSR 2-3: 19-35.

Chiba, T. and Kajiyama, J. 1941. The Vowel: its nature and structure. Tokyo: Tokyo Publishing Co.

Christensen, R., Strong, W. and Palmer, E. 1976. A comparison of three methods of extracting resonance information from predictor-coefficient coded speech. IEEE Trans. on Acoustics, Speech and Signal Processing ASSP-24: 8-14.

Dechovitz, D. 1977. Information conveyed by vowels: A confirmation. Haskins Lab. Status Report on Speech

Research SR-51/52: 213-219.

Delattre, P., Liberman, A., Cooper, F., and Gerstman, L. 1952. An experimental study of the acoustical determinants of vowel colour. Word 8: 195-210.

Essner, C. 1947. Recherches sur la structure des voyelles orales. Arch. Neerland. Phon. Exper. 20: 40-77.

Fairbanks, G. and Grubb, P. 1961. A psychophysical investigation of vowel formants. Jour. Speech and Hearing Res. 4: 203-219.

Fant, G. 1959. Acoustic analysis and synthesis of speech with applications to Swedish. Ericsson Technics 1: 1-108.
reprinted in Fant, G. 1972. Speech Sounds and Features. Cambridge, Mass.:MIT Press.

Flanagan, J.L. 1955b. A difference limen for vowel formant frequency. Jour. Acoust. Soc. Am. 27: 613-617.

Flanagan, J.L. 1972. Speech Analysis, Synthesis and Perception 2nd ed. Springer-Verlag: Berlin.

Fry, D.B., Abramson, A., Eimas, P., and Liberman, A. 1962. The identification and discrimination of synthetic vowels. Language and Speech 5: 171-189.

Fujisaki, H. and Kawashima, T. 1967. The roles of pitch and higher formants in the perception of vowels. 6th Cong. Phon. Sci.: 347-350.

Gay, T. 1970. A perceptual study of American English diphthongs. Language and Speech 13: 65-88.

Gerstman, L. 1968. Classification of self-normalized vowels. IEEE Trans. Audio. Electroacoust A-U 16: 78-80.

Helson, H. 1948. Adaptation-level as a basis for a quantitative theory of frames of reference. Psychol. Rev. 55: 297-313.

Hindle, D. 1978. Approaches to vowel normalization in the study of natural speech. In Sankoff, D. (ed.) Language Variation: models and methods. New York: Academic Press.

Joos, M. 1948. Acoustic Phonetics. Language (supplement) 24: 1-136.

Kahn, D. 1978. On the identifiability of isolated vowels. UCLA Working Papers in Phonetics 41: 26-31.

Koopmans-van Beinum, F. 1973. Formant frequencies and duration in running speech. Proc. Inst. Phon. Sci. Univ. Amsterdam 3: 49-61.

Ladefoged, P. 1967. Three Areas of Experimental Phonetics. London: Oxford Univ. Press.

Ladefoged, P. and Broadbent, D. 1957. Information conveyed by vowels. Jour. Acoust. Soc. Am. 29: 98-104.

Laver, J. 1965. Variability in vowel perception. Language and Speech 8: 95-121.

Lehiste, I. and Meltzer, D. 1973. Vowel and speaker identification in natural and synthetic speech. Language and Speech 16: 356-364.

Lieberman, P. 1973. On the evolution of human language: A unified view. Cognition 2: 59-94.

Lieberman, P., Crelin, E.S. and Klatt, D.H. 1972. Phonetic ability and the related anatomy of the newborn, adult human, Neandrathal man, and the chimpanzee. Amer. Anthrop. 74: 287-307.

Lindblom, B. 1963. Spectrographic study of vowel reduction. Jour. Acoust. Soc. Am. 35: 1773-1781.

Lindblom, B., and Studdert-Kennedy, M. 1967. On the role of formant transitions in vowel recognition. Jour. Acoust. Soc. Am. 42: 830-843.

Markel, J. and Gray, A. 1976. Linear Prediction of Speech. Berlin: Springer-Verlag.

Mermelstein, P. 1978. Difference limens for formant frequencies of steady-state and consonant-bound vowels. Jour. Acoust. Soc. Am. 63: 572-580.

Mermelstein, P., Liberman, A., and Fowler, A. 1978. Perception of vowel duration in consonantal context and its application to vowel identification. Haskins Lab. Status Report on Speech Res. SR-55/56: 123-132.

Miller, R. 1953. Auditory tests with synthetic vowels. Jour. Acoust. Soc. Am. 25: 114-121.

Nearey, T. 1977. Phonetic feature systems for vowels. Phd. thesis. reproduced by the Indiana University Linguistics Club, 1978.

Nearey, T. and Hogan, J. unpublished manuscript.

Nearey, T., Hogan, J., and Rozsypal, A. 1979. Speech signals, cues and features. In Prideaux, G. (ed.) Perspectives in Experimental Linguistics. pp. 73-93. Amsterdam: J. Benjamin.

Nearey, T., Hogan, J. and Rozsypal, A. In preparation.

Techniques in the study of English initial consonants.

Nie, N., Hull, C., Jenkins, J., Steinbrenner, K., and Bent, D. 1975. Statistical Package for the Social Sciences (SPSS) 2nd ed. New York: McGraw-Hill.

Noll, A. 1964. Short-time spectrum and "cepstrum" techniques for vocal-pitch detection. Jour. Acoust. Soc. Am. 36: 296-302.

Peterson, G., and Barney, H. 1952. Control methods used in a study of vowels. Jour. Acoust. Soc. Am. 42: 175-184.

Peterson, G., and Coxe, G. 1953. The vowels / e / and / o / in American speech. Quart. Jour. Speech 39: 33-41.

Peterson, G., and Lehiste, I. 1960. Duration of syllable nuclei in English. Jour. Acoust. Soc. Am. 32: 693-703.

Pols, L., Tromp, H., and Plomp, R. 1972. Frequency analysis of Dutch vowels from 50 male speakers. Jour. Acoust. Soc. Am. 53: 1093-1101.

Potter, R., and Peterson, G. 1948. The representation of vowels and their movements. Jour. Acoust. Soc. Am. 22: 528-535.

Potter, R., and Steinberg, J. 1950. Toward the specification of speech. Jour. Acoust. Soc. Am. 22: 807-820.

Potter, R., Kopp, G., and Green, H. 1947. Visible Speech New York: D. van Nostrand.

Shankweiler, D., Strange, W., and Verbrugge, R. 1977. Speech and the problem of perceptual constancy. In Shaw, R., and Bransford, J. (eds.) Perceiving, Acting, and Knowing: Towards an Ecological Psychology. pp. 315-345. Hillsdale, N. J.:Lawrence Erlbaum Associates.

Stevens, K. and House, A. 1963. Perturbation of vowel articulations by consonant context: An acoustical study. Jour. Speech and Hearing Res. 6: 111-128.

Stevenson, D., and Stevens, R. 1978a. A programming system for psychoacoustic experimentation. Paper presented at the 11th DECUS Canada symposium in Ottawa.

Stevenson, D., and Stevens, R. 1978b. The Alligator Reference Manual. unpublished manuscript.

Strange, W., Verbrugge R., Shankweiler, D., and Edman, T. 1976. Consonant context specifies vowel identity. Jour. Acoust. Soc. Am. 60: 213-224.

Tatsuoka, M. 1970. Selected Topics in Advanced Statistics: An Elementary Approach. No 6: Discriminant Analysis. Champaign: Institute for Personality and Ability Testing.

Tiffany, W. 1953. Vowel recognition as a function of duration, frequency modulation and phonetic context. Jour. Speech and Hearing Dis. 18: 289-301.

Tiffany, W. 1959. Nonrandom sources of variation in vowel quality. Jour. Speech and Hearing Res. 2: 305-317.

Verbrugge, R., Strange, W., Shankweiler, W., and Edman, T. 1976. What information enables a listener to map a talker's vowel space? Jour. Acoust. Soc. Am. 60: 198-212.

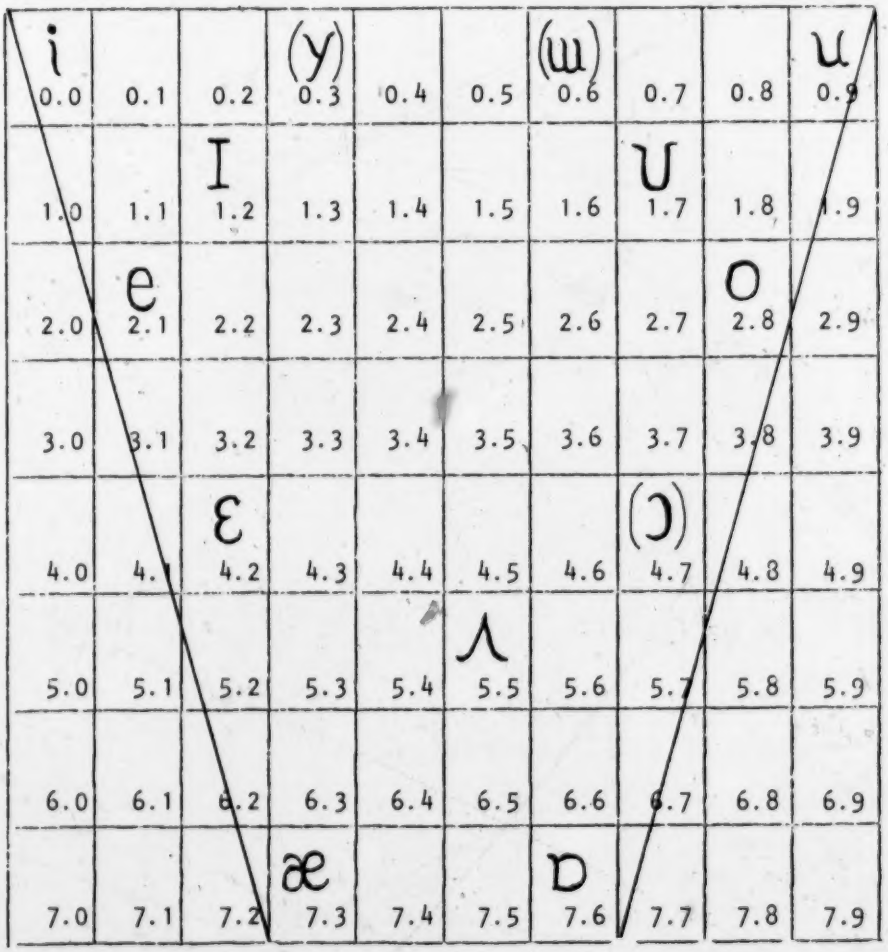
Appendix 1. Background information questionnaire

Age_____ Sex_____

1. How long have you lived in Edmonton? Approximately _____ years.
2. Where did you live before that? For how long? (List only places where you lived for more than two years.)_____

   _____

   _____
3. Is another language besides English spoken in your home? What language is it? _____

   _____
4. Do you have any hearing problems? Yes_____ No_____

Appendix 2. Vowel diagram used in phonetic transcription task